A Study of Modern Methodologies for mining sequential patterns

[1]Vaishali Shastri, [2]Mr. Sachin Mahajan

***ABSTRACT*: In this paper, we present an overview of modern sequential pattern mining techniques using data mining algorithms. Sequential pattern mining in data mining takes a lot of data base scans. Therefore it is a computationally expensive task. So still there is a need to update and enhance the existing sequential pattern mining techniques so that we can get the more efficient methods for the same task. In this paper, a study of all the modern and most popular sequential pattern mining technique is performed.**

Introduction:

The use of data mining [1,2] is placed in various decisions making task, using the analysis of the different properties and similarity in the different properties can help to make decisions for the different applications. Among them the prediction is one of the most essential applications of the data mining and machine learning. This work is dedicated to investigate about the decision making task using the data mining algorithms. Therefore an application of heart disease is reported for providing the fruitful results from the algorithms.

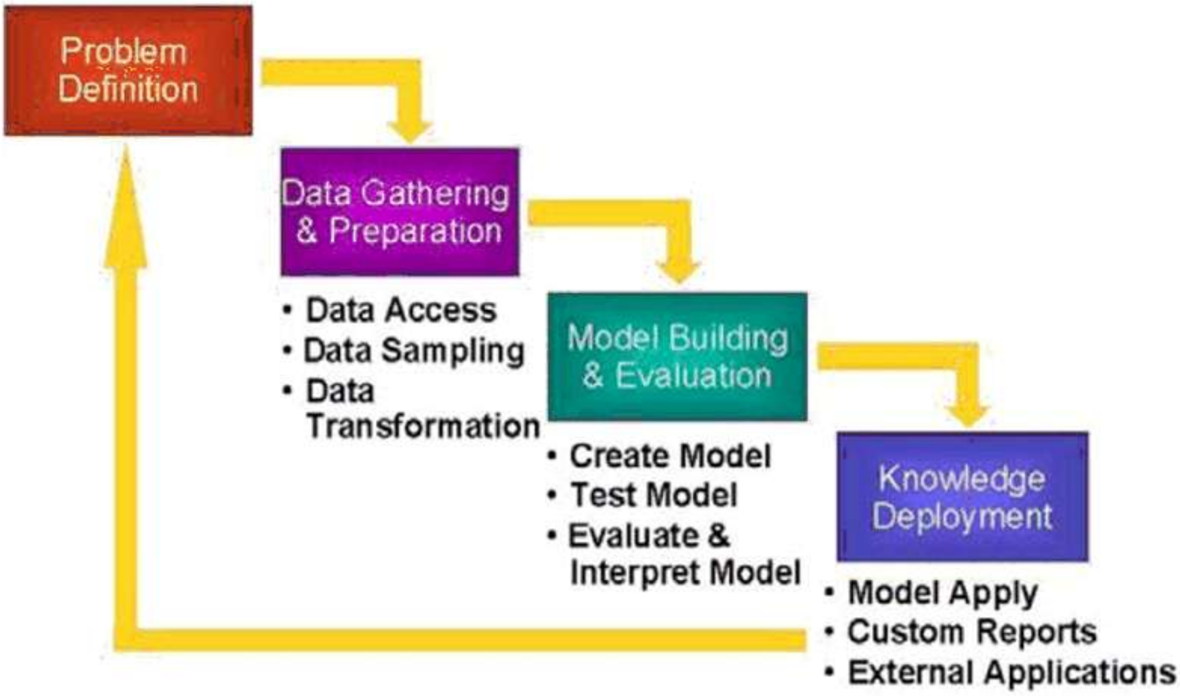


Figure 1: Data Mining

Finally, it enables them to "drill down" into summary information to view detail transactional data. With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

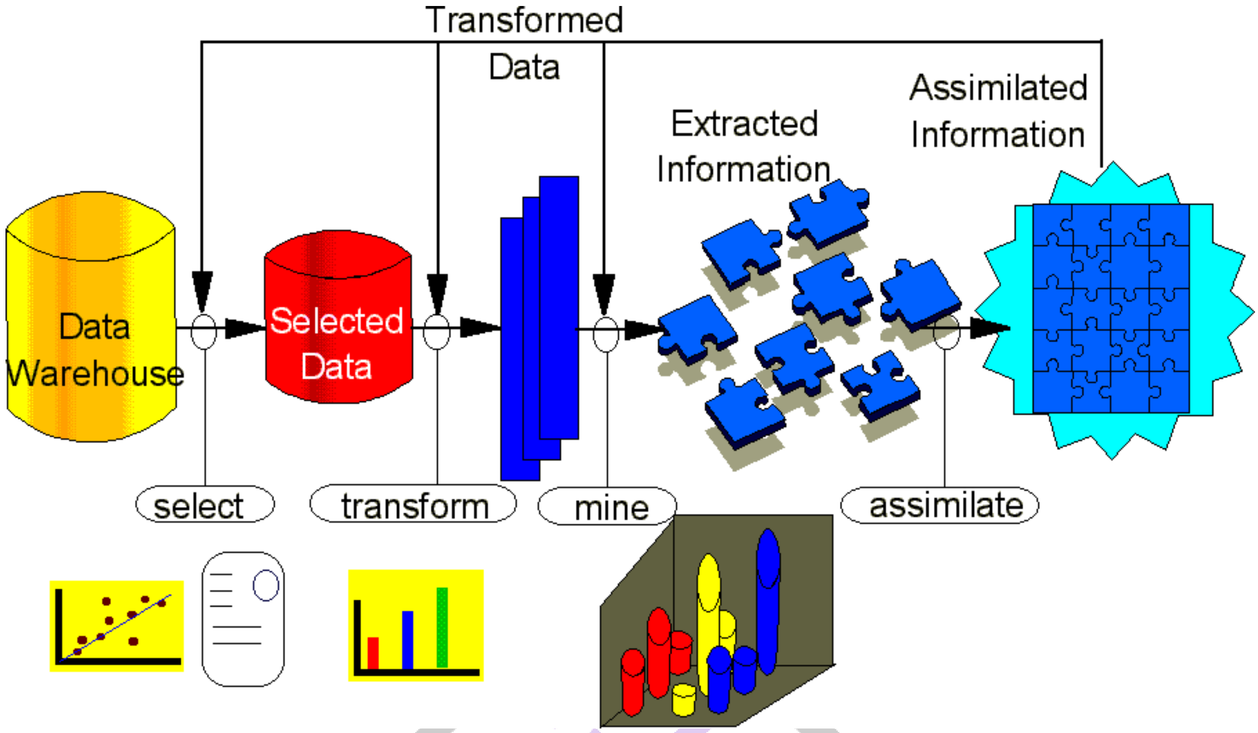


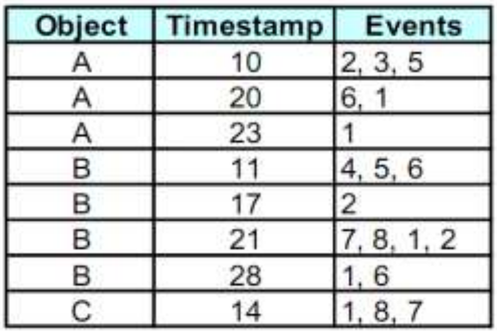
Figure 2: key steps in data mining

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. The data mining algorithms supports both kinds of learning supervised and unsupervised. In unsupervised learning only the data is used for performing the learning and in supervised technique the data and the class labels both are required to perform the accurate training. In supervised learning the accuracy [5,6] is maintained by creating the feedbacks form the class labels and enhance the classification performance by reducing the error factors from the learning model.

Literature Survey:

Frequent itemsets[1,2] and association rules focus on transactions and the items that appear there. Databases of transactions usually have a temporal information. Sequential pattern or sequential rules exploit this temporal information.
Example data:

- Market basket transactions
- Web server logs
- Tweets
- Workflow production logs



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: A Sequence Data Base

Formal Definition of a Sequence
A sequence is an ordered list of elements (transactions). Each element contains a collection of events (items). Each element is attributed to a specific time or location. Length of a sequence, |s|, is given by the number of elements of the sequence
A **sequential rule is an implication of the form** 1=>2 .It has following two associated terms:

- **Support (1=>2) = F(1 => 2)/(N)**

  Where F(1=>2) is the number of transactions in which 2 comes after 1. N is the total number of transactions.
- **Confidence (1=> 2) = F(1 => 2)/ F(1)**

Where F(1=>2) is the number of transactions in which 2 comes after 1. F(1) is the number of transactions containing 1.
Sequential Rule Mining finds all rules whose support and confidence is greater than the minimum support threshold and minimum confidence threshold respectively 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.

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. This paper presented the study of modern sequential frequent pattern mining techniques. It will help future researchers of the same area up to a good extent.

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