Frequent Itemset Mining in Distributed Data Stream Environment

Pallavi R. Kapse **Bharti W. Thakre** **Prof. S.Y. Thakur**

***Abstract*—During the study of Association Rule Mining in Distributed System it is observed that the distributed systems considered so far for the association rule mining are distributed databases but, in distributed environment like www the data sources are data streams and continuous data flow obtained from these data sources are very difficult to analysis as well as rule generation. Under this study we considered the data obtained from distributed data streams. The issues and the requirements are considered to propose the association rule mining method for a distributed stream.**

Keywords: Association rules, apriori algorithm, parallel and distributed data environment.

I. INTRODUCTION

Data mining is used to extract important knowledge from large datasets, but sometimes these datasets are split among various parts. Association rule mining is one of the data mining techniques used in distributed databases. Association rule mining is an important component of data mining. Association rule mining is a way to find interesting associations among different large sets of data item. Apriori is the best known algorithm to mine the association rules. Association rule mining, one of the most important and well researched techniques of data mining. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. distributed memory in which each processor has a private memory; and shared memory in which all processors access common memory [5]. Shared memory architecture has many desirable properties. Each processor has direct and equal access to all memory in the system. The most prominent example of a distributed environment is the Internet, where increasingly more databases and data streams appear that deal with several areas, such as meteorology, oceanography, economy and others. In addition the Internet constitutes the communication medium for geographically distributed information systems, as for example the earth observing system of NASA (eos.gsfc.nasa.gov). Other examples of distributed environments that have been developed in the last few years are sensor networks for process monitoring and grids where a large number of computing and storage units are interconnected over a high-speed network. Data Stream is the transfer of the data at a steady high speed rate sufficient to support such application as high definition television (HDTV) or the continuous backup copying to a storage medium of the data flow within a computer. Data stream requires some combination of bandwidth sufficiency and, for real-time human and perception of the data, the ability to make sure that enough data is being continuously received without anynoticible time lag. Association rule mining finds frequent itemsets which are satisfying minimum support threshold value, base on that strong association rules is generated. The association rules generate set of rule which satisfy user defined threshold value and Based on that one can develop marketing strategies. Frequent closed itemsets provide complete and condensed information for non-redundant association rules generation. Recently, much research has been done on closed itemsets mining [9, 11-13], but it is mainly for traditional databases where multiple scans are needed, and whenever new transactions arrive, additional scans must be performed on the updated transaction database; therefore, they are not suitable for data stream mining. A data stream is an ordered sequence of transactions that arrives in a timely order. Different from data in traditional static databases, data streams have the following characteristics. First, they are continuous, unbounded, and usually come with high speed. Second, the volume of data streams is large and usually with an open end. Third, the data distribution in streams usually changes with time.

II. RELATED WORK

Distributed mining methodology

Proposed methodology of building distributed data mining model can be divided into three step.

1. Choosing and implementing a data mining algorithm.
2. Selecting a data mining model quality measure.
3. Working out combining strategy and its quality measure.

In case of classification quality evaluation can be done by testing local and global model. Verifying combining strategy quality is very important step because it allows answering a basic question: are created global models good enough to meet analyst expectation?

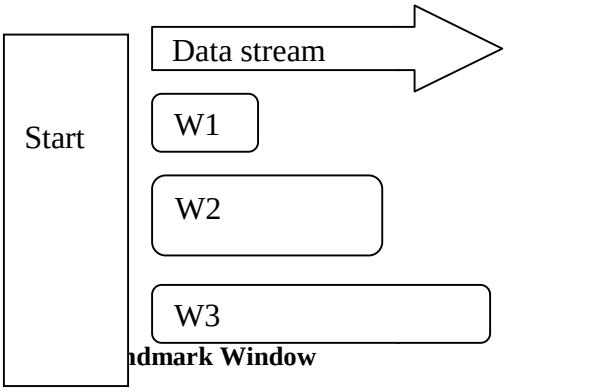
III. GENERAL ISSUE IN DATA STREAM MINING

There are some crucial issues that need to be taken into account when developing association rule for stream data.

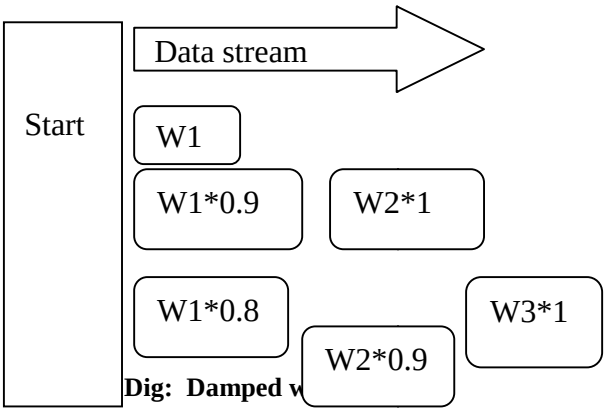
- Data Processing Model

According to the research of Zhu and Shasha[30], there are three data stream processing models, Landmark,

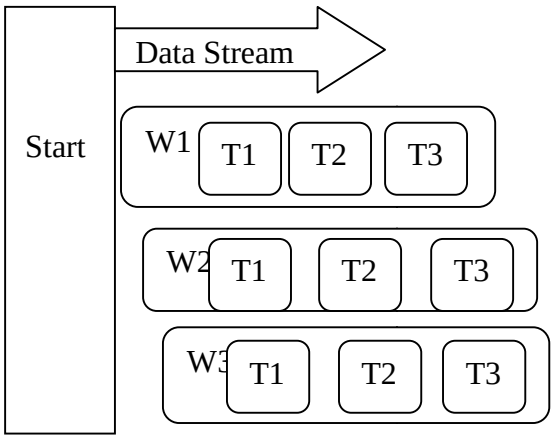
Damped and Sliding windows[11].



ıdmark Window



Dig: Damped w



The Landmark model mines all frequent itemsets over the entire history of stream data from a specific time point called landmark to the present. In this model, we treat each time point after the starting point equally important. This model is not suitable for mining where most recent information and real time data are very important such as stock market. The Damped model mines frequent itemsets over stream data. In stream data, each transaction has weight and this weight decreases with time. So in this model new and old transaction has different weights. Due to above characteristic of damped model, It is known as Time Fading model. The Sliding window model mines frequent itemset over stream data by temporary storing part of the data and processed. In this model, size of sliding window decided by need of application and system resources. Besides above mention windows, Jiawei Han et. al. proposed tilted time window model.

• Memory Management

This is major issue in mining stream data. This includes choosing of efficient and compact data structure algorithm which cans efficiently stored, updated and retrieved data. In traditional algorithm, we do multiple scan over available data. This is not possible in data stream because there is not enough memory space to store all the transaction and their counts. In simple terms, memory size is bounded and Hugh amount of data are arrives continuously. If we store the information in disks, the additional I/O operation will increase the processing time.

• Data Preprocessing

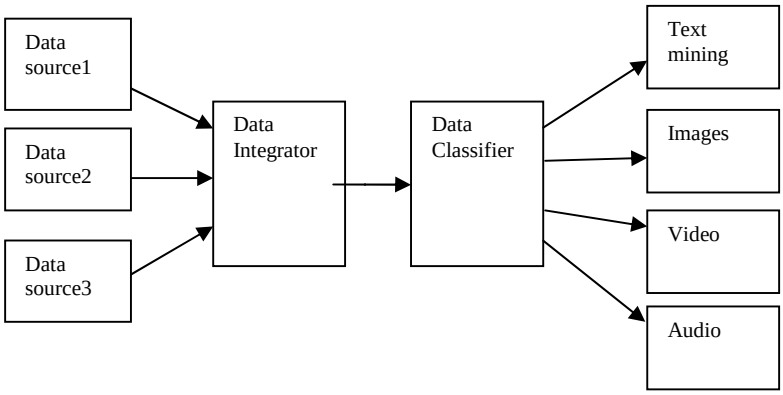
Data preprocessing is crucial aspect in the process of data mining. If data input to algorithm is not in proper format then it cannot process efficiently. So preprocessing is needed and in which existing data transform into new data which is in proper format and suitable for processing. Different data mining tools available in the market have different formats for input which makes the user forced to transform the existing input dataset into the new format.

IV. PROBLEM DEFINITION

During the study of Frequent Itemset mining in Distributed Data Stream Environment we can find out how to work data streams in association rule mining on distributed environment. In distributed environment the data sources are data streams and continuous data flow obtained from this data source is very difficult for analysis. Under this study we considered the data obtained from distributed data streams. The issues and the requirements are considered topropose the association rule mining method for a distributed stream.

V. PROPOSED SYSTEM

When Client sending a request to the Server our system mining the data before request sending to the server. There are many key challenges in data streaming mining that need to be overcome like storage, high speed processing, immediate response etc. As shown in figure data stream generated from many data sources, enters at high speed in Data stream management system (DSMS).
Following architecture to shown how to mining data in distributed environment



Dig:Architecture of stream data mining

Following figure shows that the Architecture of stream data mining. There are three types of data source that is, Data source1, Data source2, Data source3. The number of data source sending a data to the data integrator, data are integrated When data are integrated then data classifier data are generated. Data classifier involve data in there own format. Data classifier mining the data in there format. Data classifier are Text mining, Image mining, Video mining and audio mining.

There are four types of Stream data mining that is

1. Text mining
2. Image mining
3. Video mining
4. Audio mining

The Text mining, Image mining, Video mining and Audio mining are pattern generated or evaluate. When pattern are evaluate that is Text mining, image mining, video mining and audio mining are there Knowledge represented.

CONCLUSION

In this paper we discussed the Frequent Itemed mining in Distributed Data Stream Environment. We also discussed the issues that need to be considered when designing a stream data association rule mining technique. We can conclude that our paper is very efficient for stream data mining. Many papers are available on distributed data mining but in our paper we discussed stream data mining. Under this study we considered the data obtained from distributed data streams. The issues and the requirements are considered to propose the association rule mining method for a distributed stream.

REFERENCES

[1] Md. GolamKaosar, ZhuojiaXu and Xun Yi: School of Engineering and Science, Victoria University, Australia Victoria University PO Box 14428, Victoria 8001, Australia

[2] Dr (Mrs). Sujni Paul, Associate Professor, Department of Computer Applications, Karunya University, Coimbatore 641114 , Tamil Nadu, India

[3] R. Agrawal and R. Srikant , "Fast Algorithms for Mining Association Rules in Large Database," Conf. Very Large Databases (VLDB 94), Morgan

[4] Agrawal and J.C. Shafer (1996):" Parallel Mining of Association Rules", Knowledge and Data Engineering, IEEE Transactions on Volume 8, Issue 6, pp. 962-969.

[5] Cheung, J. Han, V. T. Ng,A. W. Fu, Y. Fu (1996): A fast distributed algorithm for mining association rules, 4th International Conference on Parallel and Distributed Information Systems, 18-20 pp. 31-42.

[6]Le Gruenwald, School of Computer Science. The University of Oklahoma Norman, OK 73019:CFI-Stream: Mining Closed Frequent Itemsets in Data Streams.

[7] Nan Jiang and Le Gruenwald:" Research Issues in Data Stream Association Rule Mining",The University of Oklahoma, School of Computer Science, Norman, OK 73019, USA.

[8] Y. VenkataRaghavarao * L. S. S Reddy A. Govardhan Research Scholar, JNTUH, Vice chancellor, K L University Director School of IT, JNTUH India India . Volume 4, Issue 7, July 2014 .

[9] Ms.Manali Rajeev Raut: "Association Rule Mining in Horizontally Distributed Databases" ch CSE (III sem)Dept. of Computer Science and Engineering
G.H. Raisoni Institute of Engineering and Technology for Women, Nagpur RTMNU, Nagpur

[10] Mr.NeerajRaheja: " Optimization of Association Rule Learning in Distributed Database using Clustering Technique"Computer Science Department, Assistant Professor, Maharishi Markandeshwar Engineering College/Maharishi Markandeshwar University, Mullana, Ambala, India.

[11] VinayaSawant : A Survey of Distributed Association Rule Mining Algorithms
1Asstt Prof., Department of Information Technology, DJSCE, Mumbai University, Journal of Emerging Trends in Computing and Information Sciences ©2009-2014 CIS Journal.

AUTHOR'S PROFILE

Pallavi R. Kapsereceived her BSc in Computer Science from SGB Amravati University. Presently student of MSc in Computer Science from HVPM an autonomous college.

Bharti W. Thakerreceived her BSc in Computer science from SGB Amravati University. Presently student of MSc in Computer Science from HVPM an autonomous college.

Prof. S. Y. Thakur
Assistant prof in PG Department of computer science and technology in HVPM auto autonomous college.