2nd International Conference on Emerging Trends in Engineering, Technology, Science and Management
Lakshmi Narain College of Technology, Bhopal
20th May 2017, www.conferenceworld.in
(ICETETSM-17)
ISBN: 978-93-86171-45-0

PROCESS OF DATA MINING BY USING OPTIMIZED PARTITION CONCEPT

Shruti Chandrakar[1], Shrikant Tiwari[2], Namrata Chandrakar [3]

[1,2,3] *Department of Computer Science and Engineering*
Shri Shankaracharya Technical Campus,
SSGI (Faculty of Engineering and Technology) – Bhilai (Chhattisgarh)

ABSTRACT

There are a lot of data stored electronically, in the form of raw data. In this raw data some of data is very important depends on the organization. These raw data cannot be processed directly. To process these raw data there is need to define some logic or procedure or algorithm. Before process the data and to apply the algorithm also some phases are required to filter the data like data selection and preprocessing, data transformation etc. If the size of data is in huge amount and searching are performed with some specific key many times. In this case based on the key the partition can be created which helps to search the related data only in partition dataset, but not in overall database or in warehouse. This process helps to find the result quickly.

Keywords: Data Mining, Partition of Dataset, Optimized Partition, Data Analysis.

I. INTRODUCTION

Every organization wants to mine and analyze their data in such a way that it the process used less resources and minimum response time. Also it will produce an information and pattern in such a way, that it helps the organization to increase the revenue and help to plan for market strategy. The data mining steps and techniques are also known as knowledge discovery or pattern evaluation or data extraction from database [3]. By some of the researcher the data mining technique is also known as step to knowledge discovery from database [3]. The data mining process or knowledge discovery involves the steps like data selection, data cleaning, preprocessing, data integration, transformation, data mining, pattern evaluation and the representation of knowledge [2].

Choosing a proper data mining techniques is very important. The selected technique depends on the transformed data. Every data mining techniques are not suitable for every types of data. There may be also chance to select more than one techniques to mine and analyze the data depends on the data and the requirements. The data mining techniques are also known as data mining tasks. The data mining tasks are categorized as predictive data mining and descriptive data mining.

Data mining principle, processes and techniques are not limited to special type of data; its area is very broad. The process all application for all types data. Only the difference is that, the techniques, process and algorithm may be

(ICETETSM-17)
ISBN: 978-93-86171-45-0

differing from data to data. The data can be taken from relational database, object oriented database, data warehouse, time series database, structured and unstructured database, spatial database, multimedia database and many other types of data.

1.1 Data Partition Processing

The data mining algorithms that are being developed are based on the partitioned techniques to analyze the data. The partitioned dataset is created based on the parameter. This process will helps to improve the speed of analyzing the data and also efficiency and throughput will be better while mining the data. This algorithm is first look for the partition created in previous process. If not created then it will. The partition is created, based on the frequent search or based on the keyword. Every partition is saved with its unique identification to improve the search efficiency.

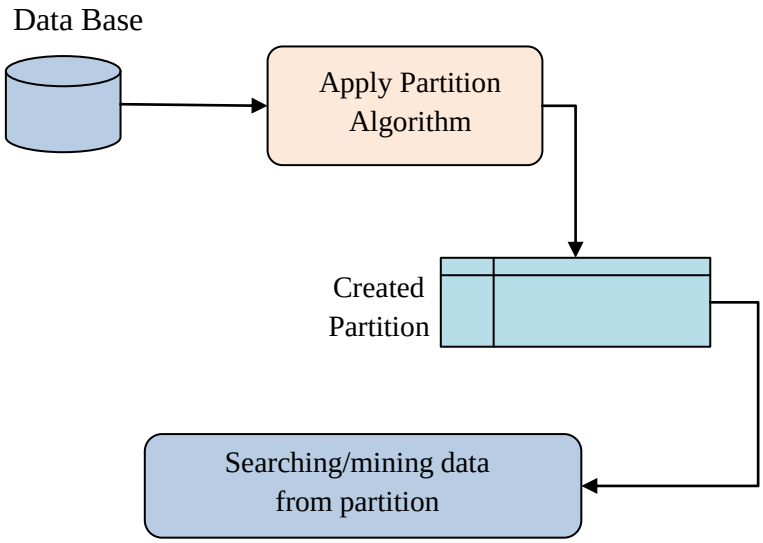

Figure 1: Flow of complete working partition process.

II. LITERATURE SURVEY

There are many techniques and algorithms are given by different researcher to mine the data in efficient way. Some of them are briefly described below.

To group the related data into cluster based on their distance and key values, is termed as a clustering of data. To handle the scalability from of mining the is take a part of all data and apply the clustering algorithm. Because the large dataset is typical to handle so if take the part of database which is relevant to mining, then it must improve the efficiency of an process. The nested partition it is important to reduce the backtracking which comes while incorrect data are partition moves. The nested partition clustering process a partitioned sub-regions that are produced with values which can have a attribute. The nested partition cluster guaranteed for scalability. The nested partition based mining algorithm select the most promising region that has the best performance of solution samples [5].

ISBN: 978-93-86171-45-0

The authors proposed techniques to partitioning clustering algorithms for detecting outliers in data streams [4]. The outlier detection process supports the clustering techniques which proved a surprised result to data miner. This process helps the miner to easily detect the ungrouped data from the clustering. BIRCH and CURE K-Means clustering are used to finding the outliers for the data stream and for comparison. The author compares the result of the algorithms based on the throughput of overall process and outlier detection. The result of this research in to CURE with K-Means clustering is performing better than the BIRCH with the K-Means. He described the tabular and graphical representation of the result to compare.

Clustering is an very important research area for data mining and processing. It creates a group of data based on their types and parameters. These data groups belong to similar type of data where one group of data is totally different from other groups. Authors' presents a paper on partition based clustering algorithm to mine a data in efficient way. In these approach partition method first creates an set of K cluster after that it use an iterative approach to add the data in to cluster based on their properties. It will improve the partition by moving the objects from one group to another group. These creates a one level of partition based on algorithms [6].

III PROPOSED TECHNIQUE

If the mining process applied for same data set for same types of searching and the data set is very large then there is problem associated with the efficiency of mining, because the process take every time same data set which is very large for same type of searching and the process is also return same type of results, so there must have a solution for this type of mining.

To overcome this problem the partitioned based mining techniques are introduced here. The basic process behind this algorithm is to creating a partition based on the search or based on the mining parameters. This will helps the user to get the result in minimum time, so increasing the efficiency of mining process. As mentioned that, the partition is created based on the searched parameter, so there is need to create and intermediate process which helps the partition algorithm to create a partition of searched keyword.

Pseudo Code of Partition Algorithm and Searching Data

```
Get the input parameter.
Check for partition
If partition created
        Then
                Do nothing, only traverse and extract the data from partition data set.
Else
        Create a partition based on parameter.
        Traverse and extract the data from partition data set.
End.
```

(ICETETSM-17)
ISBN: 978-93-86171-45-0

Flowchart of Pseudo Code of Partitioning

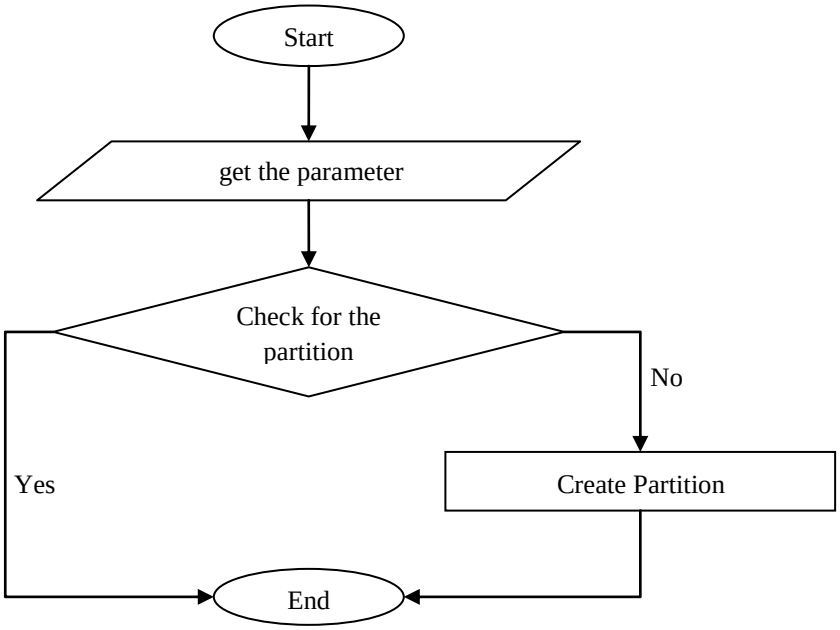

Figure 2: Flowchart of creating partition

Algorithm of Partition Dataset and Searching

The entire steps of creating a partition and searching the data are as follows. It describes the entire flow of an algorithm.

x = get the input
R=a two dimensional array to store the extracted records
Check for partition x
If x = exist
    Do nothing
Else
    Get all related values from raw database
    Create a partition dataset for that input
Traverse the partition and extract the data for given input.
Loop for k=0 ; k<(size Of Partitioned Data Set) ; k++
    if item[k]==x
        then
            R = add k^{th} values of partition to R
    else

(ICETETSM-17)
ISBN: 978-93-86171-45-0

do nothing;

End.

The above algorithm describes the entire flow of the working of the partition algorithm to search the input values. Here x is input parameter which the user wants to search. In first process the input parameter is checked that is there any partition created for that process or not. If not created, then the intermediate process will create the partition. If already created the process directly jumped into the searching phase of the algorithm

IV CONCLUSION

The ultimate aim to data mining is to provide an efficient way to get the information and analyze the business growth to increase the revenue and planning for market strategy. The next is to the mining process must be take a minimum time, efficient and high throughput. Here the applied algorithm will help the user to search their queries in minimum time, because of it use the partitioned dataset.

REFERENCES

[1] J. Zhou, L. Hu, F. Wang, H. Lu and K. Zhao, "An Efficient Multidimensional Fusion Algorithm for IoT Data Based on Partitioning", *TSINGHUA SCIENCE AND TECHNOLOGY (ISSN- 1007-0214)*, vol. 18, no. 4, pp. 369-378, 2013.

[2] U. DBD, "KDD Process/Overview", *Www2.cs.uregina.ca*, 2016. [Online]. Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html. [Accessed: 24- Dec- 2016].

[3] J. Han, "Knowledge Discovery and Data Mining, Database Systems", *Hanj.cs.illinois.edu*, 2016. [Online]. Available: http://hanj.cs.illinois.edu/(http://hanj.cs.illinois.edu/pdf/ency99.pdf). [Accessed: 15- Dec- 2016].

[4] S. Vijayarani and P. Jothi, "Hierarchical and Partitioning Clustering Algorithms for Detecting Outliers in Data Streams", *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 4, pp. 6205-6207, 2014.

[5] J. Yang, J. Kim and W. Yu, "Improving Scalability of the Nested Partition-Based Clustering", vol. 100, pp. pp.159-164, 2017 .

[6] T. Velmurugan and T. Santhanam, "A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach", Information Technology Journal, vol. 10, no. 3, pp. 478-484, 2011.