© 2017 IJSRST | Volume 3 | Issue 7 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Issues in Real Time Knowledge Discovery through Data Stream Mining

Ashish P. Joshi[1], Dr. Biraj V. Patel[2]
[1]V. P. & R. P. T. P. Science College, V. V. Nagar, Gujarat, India
[2]G. H. Patel P.G. Department of Computer Science & Technology, V. V. Nagar, Gujarat, India

ABSTRACT

The huge data are measured by recent software or hardware which are generated rapidly and highly vary such as business transaction, telecommunication call records, stock exchange, sensor networks, web logs, and computer network traffic. The challenging task is to store, retrieve and process these data sets which are considered as stream. The data stream mining is a growing technique in the field of data mining where data are analyze, process and synthesize which comes in stream. It is used to find the hidden pattern from online records of business transaction and many fields where data are frequently changes. This paper represents the current issues with this growing technique.

Keywords : DSM(Data Stream Mining) , TDM(Traditional Data Mining)

I. INTRODUCTION

In today's scenario, the use of information system is to use and share variety of information. It may be considered as business transaction, social networking, financial transaction etc. the result of this, huge amount of data assembled for storage or processing purpose. These data sets include the hidden pattern which describes some knowledge. To find out this hidden knowledge, the data mining techniques used. The data mining technique is useful for static environment where data are not very frequent. It is suitable for structured and simple data sets as like data warehouses and relational databases. In today's world, needs the fast advance and continues development where the data is generated with huge volume, different variety and very frequent which is in form of data stream. In the stream of data, data may be in complex form like spatial and sequential, text or hypertext or multimedia. Also it can be structured, semi-structured or unstructured.

So, the traditional data mining technique is not much useful for DSM. In section 2 of this paper, discussed about the DSM, section 3 of this paper is about difference between data mining and DSM discussed and in section 4 of this paper, basic of techniques & context of stream data analysis discussed.

II. METHODS AND MATERIAL

A data stream is temporal ordered sequence of data that can have any rate of volume, variety and velocity. Data Stream Mining is the technique of extracting knowledge composition from continuous, rapid data records. The main characteristics of data stream:

- ✓ Data Storage (Volume): The important piece of data from data stream can be stored and use for processing, rest of the data thrown. The data is now more than text data. One can find data in the format of videos, audios and large images on our social 2 media channels. It requires large space to store in terabyte and petabytes. Sometimes the same data is re-analyzed with different angles and even though the original data is the same the new found intelligence creates explosion of the data.
- ✓ Data Speed (Velocity): The rate of data stream is very high. It is not possible to get same stream at every time. Before some time we thought that the data before one day is recent data. But now a day data changes rapidly, today people use social media to reply frequently. The data growth is massive where too many users changes the data very rapidly. The data movement is now almost real time and the update window has reduced to fractions of the seconds.

- ✓ Data Form (Variety): The data of data stream can have any form. It can be in different format like json, xls, xml, csv or any database format. It is the need of the organization to arrange it and make it meaningful. The real world has data in many different formats and it is challenge we need to overcome.
- ✓ The figure-1 represents the working of proposed DSM model.

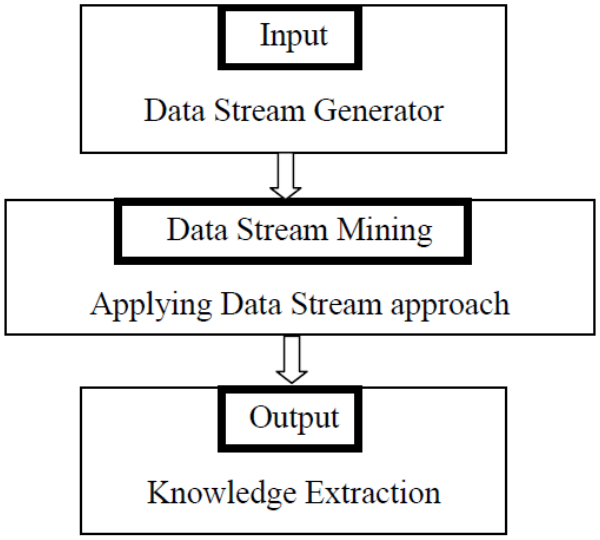

Figure 1: Proposed Model of DSM

- ✓ The model of DSM(figure 1) shows that it gathers information through data stream generator which is input that can be arrive from satellite, network sensor, business transaction, www, etc. than applying some meaningful approach or algorithm to pre-process and process arrival stream that can be incremental learning for knowledge extraction. Finally the hidden pattern outcome with meaningful knowledge can be generated for decision making.

III. DIFFERENCE BETWEEN TDM & DSM

TDM algorithms are not exactly suitable for handling DSM because of the following reasons:

Time Constraint: Limited time to apply algorithm in DSM while no time constraint in TDM.

- ✓ Memory: Limited memory can be use for DSM while no limitation for memory in TDM.
- ✓ Data: Dynamic data that can be randomly change in DSM and static data in TDM.
- ✓ Scanning: Once a stream scan, it is not available for next time. So, single pass scan for DSM while multiple time scan is possible in TDM.

IV. BASIC TECHNIQUES IN CONTEXT OF DATA STREAM PROCESSING

4.1 Random Sampling : The idea of representing a large dataset by a small random sample of the data elements goes back to the end of the nineteenth century and has led to the development of a large body of survey sampling techniques. Sampling is the process of statistically selecting the elements of the incoming stream that would be analyzed [10]. The unknown dataset size is the key problem in the perspective of data stream analysis. Special analysis using sampling depends on relationship between three parameter; data rate, sampling rate and error bounds. Designing sampling-based algorithms that can produce approximate answers that are provably close to the exact answer is an important and active area of research [2].

4.2 Histograms approximate the data in one or more attributes of a relation by grouping attribute values into subsets and approximating true attribute values and their frequencies in the data based on a summary statistics maintained in each subset. The objective of a histogram construction algorithm is to find a histogram with at most subset which minimizes a suitable function of the errors. One of the most common error measures used in histogram construction is known as the V-Optimal measure. [17] Rough data distribution can be achieved through histograms. Histograms and related synopsis structures have been successful in a wide variety of popular database application including approximate querying, similarity searching and data mining.

4.3 Sliding Window : It is considered as an advanced technique for producing approximate answers to a data stream query. The idea behind sliding window is to perform detailed analysis over the most recent data items and over summarized versions of the old ones. This idea has been adopted in many techniques in the undergoing comprehensive data stream mining system. Imposing sliding windows on data streams is a natural method for approximation that has several attractive properties. It is well-defined and easily understood. It is deterministic, so there is no danger that unfortunate random choices

will produce a bad approximation. Most importantly, it emphasizes recent data, which in the majority of real-world applications is more important and relevant than old data.[2]

4.4 Sketching : Sketching involves building a summary of a data stream using a small amount of memory[2]. It is the process of vertically sampling the incoming stream. Sketching has been applied in comparing different data streams and in aggregate queries[13]. Techniques based on sketching are very convenient to distributed computation over multiple streams. The major drawback of sketching is that of accuracy. Principal Component Analysis (PCA) would be a better solution if being applied in streaming applications [12].

4.5 Load Shedding : Load shedding refers to the process of eliminating a batch of subsequent elements (randomly or semantically) from being analyzed.[14] Load shedding has problem of drops parts in data stream, which is the reason not to prefer with mining. Still, it has been successfully used in sliding window aggregate queries. Load shedding should be performed and setting the sampling rate parameters p has two steps: 1. Determine effective sampling rates for the queries that will distribute error evenly among all queries. 2. Find values that achieve the desired effective sampling rates and satisfy the load equation.[15]

4.6 Synopsis Data Structures : Synopsis data structures embody the idea of small space, approximate solution to massive data set problems. Creating synopsis of data refers to the process of applying summarization techniques that are capable of summarizing the incoming stream for further analysis. Wavelet analysis, histograms, and frequency moments have been proposed as synopsis data structures.

4.7 Multi-resolution models : It is often used to converting one system or data into smaller part of the data based on the some assumption. One of the popular model used for this is:[18] Wavelets are one of the often-used techniques for providing a summary representation of the data. Wavelet packet transform (WPT) applies to analyze largescale electromagnetic problems. The computational domain is divided into smaller and manageable sub-domains, with coupling between these sub-domains is taken into account. As a result, a considerable reduction in the computational time and required computer memory storage is achieved.[16] This techniques is much useful in the context of memory saving.

V. CLASSIFICATION OF CHALLENGES FOR STREAM DATA PROCESSING

- ✓ Managing multiple, continues, rapid, time varying, ordered stream.
- ✓ Main memory computations: The arrival rate of stream data is variant over the time. It is also irregular and fluctuated. It challenges to utilize random memory at the random time for random data. To optimize the memory usage is big task. The summarization technique is very popular to get solution for such type of memory related issues.[9][19]
- ✓ Queries are often continues: Evaluated continuously as stream data arrives and answer the update over time.[19]
- ✓ Queries are often complex: It is beyond element and stream at a time processing also beyond the relational queries.[19]
- ✓ Data Pre-Processing: It is earlier process which applies before applying any algorithm of mining. Data which comes from stream may be erroneous, not clean, not structured or semi structured. This data pre-processing techniques apply for clean the data. It is very time consuming task for real time knowledge discovery. It must be design which give surety for quality will remain good. The light-weight pre processing techniques are useful for such types of challenges that can be easily integrated with data mining technique.[9]
- ✓ Data Structure: In any type of data related technique, the data structure plays important role. Poor data structure design affects to the efficiency of algorithm. It increases the processing that reduces the processing speed. There should be efficient data structure that manage the store, update and fetch data properly. The following techniques can be use for such types of issues:[9]
  - Incremental data structure
  - Novel indexing
  - Storage and querying techniques
  - Frequency count and time series algorithm.
- ✓ Multi level and multi dimensional data processing: Most stream data are at low level or multi dimensional in nature and it requires some

processing. Data processing needs of such applications that require collection of high-speed data, computing results on-the-fly, and taking actions in real-time. Although a lot of work appears in the area of DSMS (Data Stream Management System), not much has been done in multilevel secure (MLS) DSMS making the technology unsuitable for highly sensitive applications such as battlefield monitoring.[19]

- ✓ Visualization: The big challenges in data analysis and quick & efficient decision making. Visualization is a dominant way to provide facility for data analysis. The visualization of mining result arises many problems if the proper tools are not used. [9]

VI. CONCLUSION

In this paper, issues in existing data mining models and challenges for stream data mining discussed which shows scope of research and enhancement in stream data mining.

VII. REFERENCES

[1]. Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy: "Mining Data Streams: A Review" at https://www.researchgate.net/publicatio n/220416221 on 1-8-2017

[2]. Elena Ikonomovska, Suzana Loskovska, Dejan Gjorgjevik: "A SURVEY OF STREAM DATA MINING" in Eighth National Conference with International Participation - ETAI 2007.

[3]. Poonam Debnath, Santoshkumar Chobe: "A Quick Survey on Data Stream Mining" in International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 2948-2950.

[4]. Lalit S. Agrawal, Dattatraya S. Adane: "Models and Issues in Data Stream Mining" in International Journal Of Computer Science And Applications Vol. 9, No.1, Jan-Mar 2016 ISSN: 0974-1011

[5]. Bhavani Thuraisingham, Latifur Khan, Murat Kantarcioglu, Sonia Chib:"Realtime Knowledge Discovery and Dissemination for Intelligence Analysis"in Proceedings of the 42nd Hawaii International Conference on System Sciences – 2009

[6]. Vikas Kumar, Sangita Satapathy: "A Review on Algorithms for Mining Frequent Itemset Over Data Stream" in nternational Journal of Advanced Research in Computer Science and Software Engineering - Volume 3, Issue 4, April 2013 ISSN: 2277 128X

[7]. Tusharkumar Trambadiya, Praveen Bhanodia: "A Comparative study of Stream Data mining Algorithms" in International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012 - ISSN: 2277-3754.

[8]. Feng Chen, Pan Deng, Jiafu Wan, Daqiang Zhang, Athanasios V. Vasilakos, and Xiaohui Rong: "Review Article: Data Mining for the Internet of Things: Literature Review and Challenges" in Hindawi Publishing Corporation International Journal of Distributed Sensor Networks Volume 2015, Article ID 431047, available at http://dx.doi.org/10.1155/2015/431047.

[9]. MAHNOOSH KHOLGHI: "AN ANALYTICAL FRAMEWORK FOR DATA STREAM MINING TECHNIQUES ON CHALLENGES AND REQUIREMENTS"

[10]. P. Domingos and G. Hulten, A General Method for Scaling Up Machine Learning Algorithms and its Application to Clustering, Proceedings of the Eighteenth International Conference on Machine Learning, 2001, Williamstown, MA, Morgan Kaufmann.

[11]. G. S. Manku and R. Motwani. Approximate frequency counts over data streams. In Proceedings of the 28th International Conference on Very Large Data Bases, Hong Kong, China, August 2002.

[12]. H. Kargupta et al. VEDAS: A Mobile and Distributed Data Stream Mining System for Real-Time Vehicle Monitoring, Proceedings of SIAM International Conference on Data Mining, 2004.

[13]. A. Dobra, M. Garofalakis, J. Gehrke, R. Rastogi. Processing Complex Aggregate Queries Over Data Streams. In Proceedings of SIGMOD, 2002.

[14]. Y. Chi, H. Wang and P.S. Yu. Loadstar : Load Shedding in Data Stream Mining. In Proc. The 31st VLDB Conf., Trondheim, Norway, 2005, pp. 1302-1305.

[15]. Brian Babcock, Mayur Datar, Rajeev Motwani; Load Shedding Techniques for Data Stream Systems at http://wwwcs- students.stanford.edu/~datar/papers/ mpds03.pdf on date [01-08-17

[16]. S. H. Zainud-Deen, H. A. Malhat, K. H. Awadalla, H. A. Sharshar; Wavelet Packet Transform with Iterative Technique based on Method of Moments for large-scale problems. In Radio Science Conference, 2007. NRSC 2007.

[17]. Sudipto Guha, Nick Koudas, Kyuseok Shim; Approximation and Streaming Algorithms for Histogram Construction Problems at https://www.cis.upenn. edu/sudipto/mypapers/ histjour.pdf on 01-08-17

[18]. https://www.slideshare.net/pierluca. lanzi/ 18-data-streams on 10-08-17

[19]. Raman Adaikkalavan, Indrakshi Ray, and Xing Xie: "Multilevel Secure Data Stream Processing" at http://www.cs. colostate.edu/~iray/research/papers/dbsec11.pdf