International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (Volume 1, Special Issue 1, Oct 2013)

Data Stream Model-Issues, Challenges and Clustering Techniques

Vinay Kumar K[1], Srinivasan R[2], Elijah Blessing Raj Singh[3]

[1]Research Scholar (Part-time External), [2]Research Supervisor (External), 3Professor and Director, School of Computer Science and Engineering, Karunya University, Coimbatore, Tamil Nadu, India.

[1]kotte.vinaykumar@gmail.com
[2]srini0402@gmail.com
[3]director_cst@karunya.edu

***Abstract*—Applications such as satellite networks, telecommunication systems etc., are generating massive amount of continuous data streams every day. It is not possible to manage and analyze the streams with conventional database techniques and architectures. Hence, data streams require efficient database storage, retrieve, processing ability, and effective analysis methods to discover knowledge which is implicit in streams. Data stream mining which is one of the steps of Knowledge discovery process (KDD), is the extraction of useful knowledge structures from continuous chunks of data with high data rate. Since, incoming data is continuous; it is not possible to scan the record more than once. Stream mining is a hot research area unveils number of challenges unsolved by traditional systems. In this paper, we explore data streams showing how the Data stream management system (DSMS) is different from traditional Data base management system (DBMS), the issues and challenges in data stream mining, and tools for data stream mining.**

***Index Terms*— Data streams, DSMS, Data streams processing, indexing and Tools for data streams, Clustering in data streams.**

I. INTRODUCTION

In the last few years, the growth in hardware technology has made it for many companies to store huge amount of data. Data mining is a process of extracting knowledge from large databases[1].The data will arrive continuously as chunks of data with high rate [2]. The data sets that are endlessly grown overtime are known as data streams. The data mining has become a well developed field, the data streams is an ordered sequence of data items that arrive in timely fashion. And it is different from transactional databases. The data does not take transactional databases, but it arrives in continuous, rapid and temporally ordered data streams. While the capacity of the data is very large it posesses to number of computational issues.

The data streams are infinite & it can be found in many applications such as telecommunication systems, where data will arrive continuously as a stream in a dynamic fashion, processing and querying of these streams are challenging tasks. Satellite systems, online transactions, video surveillance, sensor networks and web applications data is not simply load the arriving into traditional databases and perform operations on it. There is a requirement of efficient mining techniques which can handle the streaming data.

In recent times the data are generating rapidly than even before from some sources. Due to this rapid growth of continuous stream of information has become a challenge for computing, storing and communication capabilities. There are some models and techniques are proposed to handle the above challenges [3] [6].Data mining functionalities for data streams such as classification, clustering, frequent patterns and association rules etc. which can be used for mining data streams.

The features of Data streams are

i. *The data in the data streams arrive continuously.*
ii. *The data streams are dynamic in nature needs quick response.*
iii. *The size of the data stream is unlimited.*
iv. *The processing of data streams can be done only once.*
v. *The storage capacity of data streams is limited so outline of the data can be stored.*
vi. *Several time of accessing a data streams is costly.*
vii. *The processed element of data stream can not be fetched easily unless it is stored in memory.*

Because of the above features of data streams, mining the data streams poses a confrontation to many researchers extracting knowledge from streamed data.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (Volume 1, Special Issue 1, Oct 2013)

It may be in internet traffic web data and Highway traffic monitoring, environmental monitoring, server log data, ATM transaction log data, stock market data, satellite networks and home land security etc.

The rest of the paper is organized as follows. In Section II we review the data stream management system (data stream model) comparison with traditional database management system. In Section III we discuss query processing in data streams. In Section IV Indexing in data streams. In Section V we explain the data stream techniques. In Section VI clustering techniques in data streams. In Section VII Data streams tools. In Section VIII research issues and challenges in data streams, finally summary of the paper and future research.

II. Data Stream Management Systems

The data stream Management system is a process to deal with uninterrupted data. It is similar to Database Management System [3]. The data base management is suitable for conservative databases in which the data is stored in the form of relations. The major features of DSMS is to handle the data effectively, dynamically changing behavior of data streams & approximate query processing which have a limited amount of memory. The data stream management system (DSMS) differs from DBMS is shown in table 1. Data streams management system is suitable for continuous queries, sequential access, limited memory that arrives inconsistent with high data rate.

Table. 1
Comparison between DSMS and DBMS

Data stream management system (DSMS)	Data base management system (DBMS)
Query semantics is a Continuous queries that produce results as stream tuples arrive time advances	Query semantics is an Onetime queries over data sets that do not change while a query runs
It is suitable for sequential access	It can suitable for random access
It can works with unstable data	It can works with constant data
The data in DSMS is a Continuous, rapid and time-varying data streams.	The Data in DBMS is Finite data sets on disk, along with auxiliary structures like indexes.
Queries processed by incremental operators that are scheduled based on tuple arrival.	Queries processed by scan-based or index-based operators that read data from disk.
It can has a limited main memory	It can has an unlimited secondary storage
It can extremely works with high update rate	It can relatively works with low update rate

One of the challenges for a DSMS is to handle potentially infinite data streams using a fixed amount of memory and no random access to the data. There are different approaches to limit the amount of data in one pass, which can be divided into two classes. For the one hand, there are compression techniques that try to summarize the data and for the other hand there are window techniques that try to partition the data into (finite) parts.

Data stream model: The arrival of data in dynamic, time varying, changeable and enormous. In data stream model data items may be relational tuples. For example: server log data, sensor readings, in these applications it is not possible to store on traditional database systems and work on it. The data stream systems straight forwardly support the continuous queries [3,4]. In data stream model, the input data is not stored in disk but the data stream arrives continuously as chunks of data. The data stream model is distinct from conservation relational model in the following manner

I. Data streams have an unlimited in size
II. The data element from data streams is processed then it is detached, it is difficult to retrieve unless it is stored in memory.
III. The data in the data stream is continuous.

III. Query Processing In Data Streams

While computing, the query processing in the data stream model can have its own issues & challenges limitless memory requirements. The sizes of the data streams are potentially boundless. Hence, the possibility to get the exact answer by computing a data stream is very less. The exterior memory methods for handling data sets are larger than main memory that is not suitable for data stream applications. Some of the issues are given in the following.

Estimate query response: By using the limited memory it is not possible to generate exact response for data stream queries. Histogram based methods gives estimate query response for correlated aggregate queries over data stream. Gilbert et at [5] proposed a method for small space summary of data streams which gives estimate query response for aggregate queries (Group functions).

Sliding Windows: In the sliding window method provoked by the supposition is considering the recent data from data streams rather than old data of the data stream. The processing and analysis over the recent data and summarized of the older data [6].

Many DSMS are based on the query processing in DBMS by using languages to express queries, which are translated into a plan of operators. These plans can be optimized and are executed. A query processing often consists of the following steps.

a) *Formulation of continuous queries:* The formulation of queries is mostly done using declarative languages like SQL(structured query language) in DBMS. Since there are no standardized query languages to express continuous queries, there are a lot of languages and variations. However, most of them are oriented to SQL like Continuous query languages (CQL), Stream SQL. For an example, if windows are used for the processing, the definition of the window has to be expressed.

b) In Stream SQL,For example, a query with a sliding window for the last 25 elements looks like follows:

Select AVG (price) from stream123 [SIZE 25 ADVANCE 1 TUPLES] where value > 100;

This above query calculates the average value of "price" of the last 25 tuples, but only considers those tuples for the average calculation where price is greater than 100.0.

In the next step the declarative query is translated into a logical query plan. A query plan is a directed graph where the nodes are operators and the edges describe the processing flow. Each operator within the query plan encapsulates the semantic of a specific operation like filtering or aggregation.

In DSMS that process relational data streams, the operators are equal or similar to the operators of the relational algebra so that there are operators for selection, projection, join, or set operations. This operator's concept allows flexible and versatile processing of a DSMS.

b) Optimization of queries: It is the logical query plan can be optimized, which strongly depends on the streaming model. If there are relational data streams and the logical query plan is based on relational operators from the relational algebra a query optimizer can use the algebraic equivalences to optimize the plan.

c) Transformation of queries: The logical query plan must be transformed into an executable counterpart. This is called a physical query plan. The distinction between a logical and a physical operator plan allows more than one implementation for the same logical operator.

The join, for example, is logically the same, although it can be implemented by different algorithms like a nested loop join and sort merge join and because these algorithms also strongly depend on the used stream and processing model. Finally, the query is available as a physical query plan.

d) Edit source: Since the physical query plan consists of executable algorithms, it can be directly executed. For this, the physical query plan is installed into the system. The query plan is connected to the incoming sources, which can be everything like connectors to sensors. The top of the graph is connected to the outgoing sinks, which may be for example visualization. Since most DSMS are data-driven, a query is executed by pushing the incoming data element from the source through the query plan to the sink. Each time when the data element passes an operator, the operator performs its specific operation on the data element and forwards the result to all successive operators.

IV. INDEXING IN DATA STREAMS

The data stream management system requires indexing in order to extract knowledge from continuous chunks of data.

The data stream indexing[7] is not precisely same as conservative data base management systems because of the continuous flow of data. Some of the distinct features between conventional data base management system indexing and data stream management system needs ordered indexing.

- The sizes of the data streams are unlimited with high data rates so data stream management system requires indexes to be store in the main memory.
- Conventional DBMS is static method in which once data has been filled in the data base, we can efficiently perform insert, update and delete operations. Window indexing is needed in data stream management system in order to handle insert, delete and update with high data rate.

V. DATA STREAM METHODS

Problems and challenges in data stream mining that take place from the statistical and computational methods can be solved by using the research techniques which consists of

- Investigating a small set of the entire data set.
- For efficient utilization of time and space by using the algorithms.

We classify these solutions into 2 types.

5.1 Data based methods
5.2 Task based methods

The First method refers to summarization of entire data set.

5.1 The database methods are

5.1.1 Sampling
5.1.2 Load shedding
5.1.3 Sketching
5.1.4 Synopsis data structures and aggregation**.**

These methods are explained in brief

5.1.1 Sampling

Sampling [8] uses the statistical technique for a long time which makes a probabilistic choice of data item to be processed.

Problems related to sampling technique while analyzing the data stream are

- The size of dataset is unlimited.
- Sampling does not deal with the problem in which the data rates are of fluctuating.

5.1.2 Load shedding

Load shedding is a method of dropping a small set of data streams during periods of overload [9, 10]. Load shedding is used in querying data streams for optimization. Load shedding has two steps. Firstly, for each query choose target sampling rates. In the second step, place the load shedders to realize the targets in the most efficient manner. It is difficult to use load shedding with mining algorithms because it drops the chunks of data stream that can be used in represent a pattern of interest in time series analysis**.**

5.1.3 Sketching

Sketching [3, 6] is a process of vertically sampling the incoming stream by projecting a subset of features.

5.1.4 Synopsis data structures

Creating a summary of data refers to the process of applying summarization techniques to the incoming data stream for analysis. Wavelet analysis [11], histograms, quartiles and frequency moments are proposed for synopsis data structures. Synopsis data structures could not represent the characteristics of the entire data set; it can provide approximate answers when we use the synopsis data structures.

There are a different methods are used for construction of synopsis data structures. These methods are as follows.

•Sampling methods
•Histograms
•Wavelets[12]
•Aggregation
•Micro Cluster based summarization[13,14]
•Sketches

5.2 Task based methods:

In order to address the challenges in data stream processing the task based methods uses the changing the existing method or inventing a new one. These methods are as follows.

5.2.1 Sliding window method:

In this method the user is more concerned with the current data for analyzing data streams and detailed analysis can be done on current data rather than summarized version of the older data [15].

5.2.2 Approximate algorithm

These algorithms [6] are designed for solving the hard problems and it can give approximate answer with error bounds. Due to this, many researchers are working on data streams to provide accurate solutions to problems in data stream mining. By using the available resources we can't solve the problems so we need to use some other tools to solve the problems in data stream mining.

5.2.3 Algorithm output granularity (AOG)

It is suitable to handle when the data rates are fluctuating. It gives the results by using the available memory and with time constraints [16]. There are three stages in this process (i) Mining data streams, (ii) Adaptation of resources (iii) Merging the generated structures when running out of memory. The AOG performs the local data analysis on resource constraint device that generate streams of information. The algorithms are used with association, clustering and classification.

VI. Data Streams Using Clustering

Data stream mining has attracted many researchers over the last few years. The numbers of algorithms are proposed for extracting knowledge structures from continuous of streaming data.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (Volume 1, Special Issue 1, Oct 2013)

The stream mining becomes a challenging task because of the feature of the data streams. The data mining have different functionalities like classification, clustering and association rule mining etc. In this paper we are focusing on data stream mining using clustering.

Clustering data streams: Clustering is one of the sub area in data stream mining. The data is arranged into different clusters based on similarity of data elements. Clustering can be used in many applications such as telecommunications, spatial data mining, pattern recognition, banking transactions and web log server data.

Features:

1) *Changing data overtime* – Data streams are continuous flow of data and clustering can be incremental in order to handle the eronomous amount of data with high data rate.
2) *Concept evolution* – For handling dynamic features of data streams clustering must address the new cluster and discarding the old clusters. In concept evaluation new types of class normally holds new set of features so different chunks may have different feature sets.
3) *Data fading* – Due to the massive amount of incoming data in data streams, it mainly concentrates on current data and giving less preference to previous data.
4) *Clustering approach* – Different types of clustering approaches [1] are used for mining the data streams such as hierarchical based clustering, density based clustering, grid based clustering, model based clustering and partition based clustering.
5) *Concept change* – It is the change of the underlying concept over time.

Some of the techniques for data stream clustering are as follows.

E-stream [17] is a data stream clustering techniques which carry the five types of evaluation in the streaming data.

- Arrival of new cluster
- Discarding an old cluster
- Splitting of a large cluster
- If similar types of clusters then it could to be merged
- Changing the behavior of the cluster. It uses the fading cluster technique with suitable stream data.

Charikar etal.[18] proposed a another algorithm using k-medians. It defeats the problem of increased approximation factor proposed by guha [22] with the increase the number levels used to results in the final solution of the divide and conquer approach.

HUE-stream [19,24] extends E-stream to support the uncertainty in the heterogeneous data. It supports for both numerical and categorical attributes.

O'challaghan et al.[23] proposed stream algorithm for data streams using clustering.

a) First initially it determines the size of the sample then
b) Applying a local search algorithm if the sample size is greater than pre specified equation result.
c) The same steps are applied to every chunk of data streams.
d) Applying the local search algorithm to the cluster centered generated in the previous iterations.

Aggarwal et al.[20] proposed a algorithm called Clustream. In this method clustering process is decomposed into 2 components. firstly online component, next offline component. Many experiments are conducted in order to prove the efficiency of the algorithm.

Aggarwal et al.[20,24] proposed a algorithm called HPstream for high dimensional data streams using clustering. It uses a fading cluster structure to summary of data and giving more preferences to current data than previous data.

Guha et al. [22] proposed a analysis of clustering data streams by using k-median technique. This algorithm uses a small space to make a single pass over the data streams. It requires time O(nk) and space O(ne) and k clusters where n is number of points and e<1.Firstly it starts calculating the size of the sample using clustering along with the available Memory into 2k. Secondly clusters the points for number of samples into 2k and same procedure are applied. To the remaining levels and at last clusters the 2k clusters into k clusters.

Babcock et.al[25] is improved the k-median method using Exponential data structures proposed by guha et al [22].

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (Volume 1, Special Issue 1, Oct 2013)

Table 2
Clustering techniques

S. no	Author	Approach	Features Used
1	P.Domingos	Machine learning algorithms	Hoeffding bound
2	O'challaghan	Stream & local search algorithm	Incremental learning
3	C.C.Aggarwal	HPstream	Fading cluster structure
4	Aggarwal	CLUstream	Partitioning & Hierarchical Clustering
5	Gaber	AOG	Light weight Cluster
6	S.Guha	K-medians	One pass over the data stream using a small space
7	Charikar	K-median	Increasing the number of levels of Guha algorithm
8	W.Meesuksabai	Hue-stream	Hierarchical clustering
9	K.Udommanetana kit	E-stream	Fading cluster technique
10	Babcock	K-median	Exponential Histogram Data structure

Gaber et al. [21] developed a LWC and it is a AOG based algorithm. Firstly it adjust the threshold that gives the minimum distance measure between data items in distinct clusters and it adjustments can be done based on a time frame.

Domingos et al.[26] proposed a technique for scaling up Machine learning algorithms, these algorithms are implemented and calculated using synthetic data sets and web data streams.

VII. Tools for Data Streams

These are some of the tools for data stream mining.

1. VFML: The VFML (Very Fast Machine Learning) tool used for mining high-speed data streams and very large data sets. VFML consists of three main components.

1) It is a collection of tools and APIs that help a user develop new learning algorithms.
2) This is a collection of implementations of important learning algorithms.

2. Rapid miner: Rapid Miner is unquestionably the open source system for data mining. It is available as a standalone application for data analysis.

Features

- It works on every major platform and operating systems.
- It is a open source system for data mining and analysis.

The MOA tool: MOA **(**Massive Online Analysis) is a structure for data stream mining. It consists of a collection of machine learning algorithms and tools for evaluation. *MOA* can be easily extended with new mining algorithms; it can also be used with new stream evaluation measures. The goal is to provide a standard for the stream mining society.

MAIDS: Mining alarm incidents from Data Streams [24] It find outs the alarm incidents from data streams and it is suitable for various functionalities like online stream classification, online clustering data streams and pattern mining.

VIII. Research Issues

There are lots of open research issues in the field of data stream mining. Some of the issues are given in the following.

- Preprocessing of data streams
- Reviewing the incoming data for reliability.
- Handing the continuous flow of data streams
- Real applications of stream data management and processing data
- Satisfying the user requirements
- Scalability of data stream mining systems.
- Accuracy of results while dealing with continuous flow of data.
- Evolving data and space limitations are issues in clustering data streams
- Interactive processing is required in order to change the constraint during the processing time.
- Evolving data, one time scan are issues in clustering data streams.

Challenges:

- Changing and asymmetrical arrival of data rate and alternative data arrival rate over time.
- Providing a Quality of results using data mining and computerization of preprocessing methods.

- Tiny amount of memory and huge volume of data streams.
- Limited resources for storage and computing.
- Problems are in during the data analysis and quick decision has to take by end user [27].

IX. Conclusion And Future Work

Data stream mining is an emerging field which needs to address more challenges and issues which we presented in a concise manner.In this paper we concentrated mainly on clustering techniques which would discover the patterns in the data, Next there is a need to have better data models which would handle stream data. Finally more effective algorithms still have to be developed to get accurate results in evolving data with a single scan with efficient indexing structures.

REFERENCES

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, J. Kacprzyk and L. C. Jain, Eds. Morgan Kaufmann, 2006, vol. 54, no. Second Edition.

[2] Mohamed Medar Gaber,Arkady Zaslavsky and Krishnaswamy,"Mining Data Streams,A Review", SIGMODC Record,Vol 34,No 2, 2005.

[3] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In Proceedings of PODS, 2002.

[4] Stanford Stream Data Management (STREAM) Project. http://www-db.stanford.edu/stream.

[5] A. Gilbert, V. Kotidis, S. Muthu krishnan and M. Strausss Surfing wavelets on data streas: One-pass summaries for approximate aggregate queries. In proc of the Intl. Conf on very large Data bases, pages 79-88, 2001.

[6] Mutheekrishinan. S (2003). Data streams algorithms and Applications In proceedings of the fourteenth annual ACM-SIAM symposium on discrete algorithms.

[7] Http//:www.it.uu.se/research/group/udbl/dsms ordered indexing.

[8] 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.

[9] N.Tatbul and S. Zdonik. A subset-based load shedding approach for aggregation queries over data streams.In Proc.Int. Conf on very Large Data Bases (VLDB), 2006.

[10] Babcock, M Datar, and R Motwani. Load Shedding Techniques for Data Stream Systems (short paper) In Proc. of the 2003, Workshop on Management and Proc of Data Streams,June 2003.

[11] A.C.Gilbert, Y.Kodidis, S.Muthukrishan, M.Strauss: One-Pass Wavelet Decompositions of Data Streams. TKDE (3), 2003.

[12] Keirn D. A. Heczko M. (2001) Wavelets and their Applications in Databases. ICDE Conference.

[13] Aggarwal C, Han 1., Wang 1., Yu P. (2003) A Framework for Clustering Evolving Data Streams, VLDB Conference.

[14] Zhang,T.,Ramakrishnan, R, Livny, M.: Birch : An efficient data clustering method for very large databases, In: SIGMOD,Montreal, Canada, ACM (1996).

[15] Chang, 1. H. and Lee, W. S. (2005): estWin: Online data stream stream mining of recent frequent item sets by sliding window method. Journal of Information Science, 31(2): 76-90.

[16] Gaber, M, M., Krishnaswamy, S., and Zaslavsky, A., On-board Mining of Data Streams in Sensor Networks, Accepted as a chapter in the forthcoming book Advanced Methods of Knowledge Discovery from Complex Data, (Eds.) Sanghamitra Badhyopadhyay, Ujjwal Maulik, Lawrence Holder and Diane Cook, Springer Verlag, to appear.

[17] K. Udommanetanakit, T. Rakthanmanon, and K. Waiyamai, "E-stream: Evolution-based technique for stream clustering,"in Proceedings of the 3rd international conference on Advanced DataMining and Applications, ser. ADMA '07Berlin, Heidelberg: Springer-Verlag, 2007, pp. 605–615.

[18] M.Charikar, L.O'Callaghan, and R.Panigrahy. Better streaming algorithms for clustering problems In Proc of 35th ACM Symposium on Theory of Computing, 2003.

[19] W. Meesuksabai, T. Kangkachit, and K. Waiyamai, "Hue-stream: Evolution-based clustering technique for heterogeneous data Streams with uncertainty." in ADMA(2), ser. Lecture Notes in Computer Science, vol. 7121. Springer, 2011, pp. 27–40.

[20] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for projected clustering of high dimensional data streams," in Proceedings of the Thirtieth international conference on Very large data bases – Volume 30, ser. VLDB '04. VLDB Endowment, 2004, pp. 852–863.

[21] Gaber, M, M, Zaslavsky, A, and Krishnaswamy, S., Towards an Adaptive Approach for Mining Data Streams in Resource Constrained Environments, the Proceedings of Sixth International Conference on Data Warehousing and Knowledge Discovery - Industry Track (DaWak 2004),Zaragoza, Spain, 30 August - 3 September, Lecture Notes. in Computer Science (LNCS), Springer Verlag.

[22] S.Guha, N.Mishra, R.Motwani, and L.O'challaghan clustering data streams.In Proceedings of the Annual Symposium on Foundations of Computer Science.IEEE,November 2000.

[23] Yogita and D. Toshniwal, "Clustering Techniques for streaming data," in Proceedings of IEEE,Intl. Conference on Soft Computing for Problem Solving, India, 2012.

[24] Mahnoosh kholghi,Mohammed reza keyyvanpour, An analytical frame works for data stream mining techniques based on challenges and requirements, vol.3, Mar 2011.

[25] B.Babcock, M.Datar, R.Motwani, L.O' Callaghan: Maintaining Variance and K-Medians over Data stream Windows, Proceedings of the 22nd symposium on Principles of Database Systems.

[26] P. Domingos and G.Hulten, A general Method for scaling Up Machine Learning Algorithms and its applications to Clustering, Proc. of the Eight Intr. Conf on Machine Learing, 2001, Williamstown, MA, Morgan Kaufmann.