IMPACT: International Journal of Research in Engineering & Technology (IMPACT: IJRET)
ISSN(E): 2321-8843; ISSN(P): 2347-4599
Vol. 2, Issue 5, May 2014, 143-152

IMPROVE FREQUENT PATTERN MINING IN DATA STREAM

HIMANSHU M. SHAH[1] & NAVNEET KAUR[2]

[1]Master of Technology, Department of Computer Science, Lovely Professional University, Punjab, India

[2]Assistant Professor, Department of in Computer Science, Lovely Professional University, Punjab, India

ABSTRACT

A data stream is continuous, rapid, unbounded sequence of data. Mining Frequent pattern in stream data is very challenging because data can be scan one time only. Due to this reason traditional approach cannot be use for data stream. In this paper we give overview of growing field Data Stream mining. Generating and maintaining association rule from stream data is not easy task. Association rule help retailer and business manager by providing hidden useful information. There are many algorithm proposed for efficiently mining Data stream which we discuss in this paper.

KEYWORDS: Association Rule Mining, Data Stream, Frequent Pattern Mining, Stream Mining

I. INTRODUCTION

Association rule mining finds frequent itemsets which are satisfying minimum support threshold value, base on that strong association rules is generated. The association rule generate set of rule which satisfy user defined threshold value and Based on that one can develop marketing strategies. Not only in sales marketing, there are many areas such as inventory management, sales management and strategy management etc. in which this kind of strong rule become very helpful.

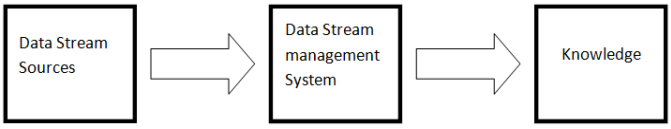

Figure 1: General Framework

Now a days, Many organization, social website, sensor network and many other sources generate hugh amount of data and they are high speed in nature. Therefore Researchers and big organization got attention toward Data streaming mining. Mining from rapid, unbounded, dynamic stream of data is very challenging. Therefore it is major research topic.

Data mining is a technique to extract hidden useful information from large database. There are many algorithms such as Apriori and FP Growth which can efficiently discover pattern and trends from database. These are traditional algorithm which cannot used in Data stream mining. Since these algorithm need to scan more than one time to generate frequent pattern from data, therefore it cannot apply because in stream data we can scan data only ones.

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). In DSMS, algorithm may use different types of model based on user interest.

Table 1: Difference between DBMS (Database Management System and DSMS (Data Stream Management System)

	DBMS	DSMS
Data type	Static data	Stream Data
Relationship	Persistent data	Volatile data stream
Access	Random	Sequential
Query	One time	Continuous
Storage	Passive repository	Active repository
Available memory	Flexible	Limited
Algorithms	Processing time is not a constraint	Processing time is most important as data may skip
Results	Accurate	Approximate
Response speed	No time requirements	Real-time requirements
Data scan	Flexible	One time scan only
Data Schema	Static	Dynamic

As shown in table 1[10][15], Stream data are continuous, rapid, time varying and unpredictable and unbounded and require quick repose. Therefore traditional DBMS and algorithms which are designed for static data are not suitable for mining stream data because it cannot fulfil the requirement of stream data mining. Example of data stream includes Sensor network, web click-stream data, computer network monitoring, telecommunication connection data, Intrusion detection, readings from sensor nets and stock quotes, Environmental and weather data. This type of data is called a data stream and dealing with data streams has become an increasingly important area of research.

This paper will focus on the following sections. In Section 2 we present motivation of data stream mining. Section 3 describe preliminaries and gives example of frequent pattern mining. Section 4 discusses various issues regarding data stream mining. Section 5 discusses analysis of frequent pattern mining papers over data stream. At the end Conclusion and future work of this paper are discussed in section 6.

II. MOTIVATIONS

Now a day's many organisation and researcher taking interest in stream data mining. Currently there are many sources produces stream of data continuously which are unbounded, rapid and highly dynamic in nature. Example include sensor networks, wireless networks, radio frequency identification (RFID), customer click streams, telephone records, multimedia data, scientific data, sets of retail chain transactions, etc. These sources are called data streams.

Data stream mining is important research topic in research community and the number of researches also growing in this field. Many algorithm and technology developed and evolved for handling complexity and volume of data, still there is need of general purpose algorithm, model with smaller space complexity, smaller time complexity and high performance in nature[15]. Since it is under research area there are wide chances of exploration.

III. PRELIMINARIES

A data stream $D=\{B_1,B_2,\ldots,B_N\}$ is a an infinite sequence of batches where each batch Bi contains a set of transactions i.e. $Bi=\{T_1,T_2,\ldots,T_K\}$ where $k > 0$. Each transaction $T=(TID,I_1,I_2,\ldots,I_n)$ is a set of items such that $T \subseteq D$, while n is called the size of transaction and TID is unique identifier of the transaction. An itemset is a non-empty set of items. An itemset with size k is called an m itemset.

There are many types of window model which use to process data stream. A window, W, can be (1) either time-based or count-based, and (2) either a landmark window or a sliding window. W is time-based if W consists of a sequence of fixed-length time units, where a variable number of transactions may arrive within each time unit. W is count-based if W is composed of a sequence of batches, where each batch consists of an equal number of transactions. W is a landmark window if W = $\{T_1, T_2, \ldots, T_T\}$; W is a sliding window if W = $\{h_{T-w+1}, \ldots., T_T\}$, where each T_i is a time unit or a batch, T_1 and T_T are the oldest and the current time unit or batch.

An association rule is an implication of the form A ⇒ B, where $A \subset W$, $B \subset W$, and $A \cap B = \emptyset$. It helps to discover combination of goods. The **occurrence frequency or frequency** of an itemset(x) I is the number of transactions that contain the itemset in a batch B and denoted as freq(x). Occurrence frequency is also called as absolute support. **Support** of X denoted by supp(X) is freq(X) / N, where N is total number of transactions received in W in data stream. It is also called as relative support. Support (A⇒B) = P (A ∪ B). **Confidence** of a rule X ⇒ Y denoted by conf(X) is supp(X ∪ Y) / supp(X) where c is the percentage of transactions received in W, containing A that also contain B. Confidence (A⇒B) = P (B | A). The lift value indicates that how many more times itemset occurred than expected. it can interpret the importance of a rule. It is measure of a rule but it cannot be define as minimum lift to minimum support or minimum confidence. Lift(X ⇒ Y) = confidence / expected confidence = Supp(X U Y) / Supp(X) * Supp(Y)

An itemset X is called as **frequent itemset (FI)** if supp(X) >= minSupp, where minSupp is user defined minimum threshold support. An itemset X is **closed** in W if there exists no proper super-itemset Y such that Y has the same support count as X in W. An itemset X is a **frequent closed itemset (FCI)** in W if X is both closed and frequent. An itemset X is a **frequent maximal itemset (FMI)** in W if X and Y are frequent, and there exists no super-itemset Y such that X ⊂Y.

To mine FIs/FMIs/FCIs over a window in data stream, it is necessary to keep infrequent itemsets, because it may become frequent later.

Example

In the example, there are 5 transactions. Each transaction contains number of items. For the simplicity we use A, B, C, D, E, F letters to denote items.

Table 2

ID	Itemset
1	A,B,C,E
2	A,D,E
3	A,B,C,D,E,F
4	B,C,E,F
5	B,C,D

In this example, A,B,C,D,E,F occurred 3,4,4,3,4,2 times respectively.

Here minimum support=4 is set. Therefore only 3 item B, C, and E satisfy minimum support threshold. Therefore this items are frequent items because their occurrence values are equal to the threshold value. Items A, D and F are called as infrequent item sets. So they are omitted. Thus this is called as frequent pattern mining.

Let's take nonempty subsets of l = {B, C, E} are {B, C}, {B, E}, {C, E}, {B, {C}, and {E}. If the minimum confidence threshold is, say, 80%, then only 2 rules are output and they are (1) { B ^ E } ⇒ {C}, Confidence=100% (2) {C ^ E} ⇒ {B} = 100%. Relative Support = 3/5 = 0.6 that means it occurs in 60% of all transactions.

Lift(X ⇒ Y) {B, E} ⇒{C} has a lift of 0.6 / (0.6 * 0.8) = 1.25.

IV. GENERAL ISSUES 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].

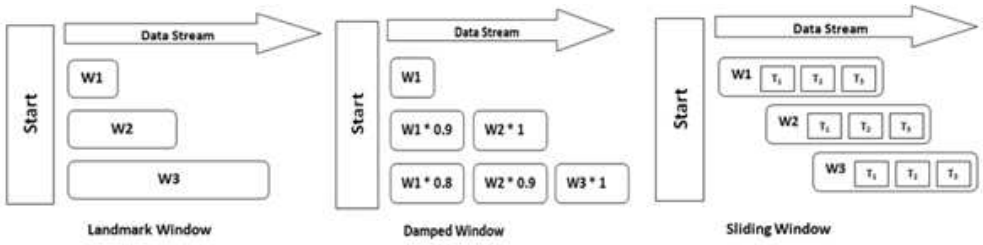

Figure 2

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. In this model, we are interested in frequent itemsets over a set of windows[29]. Each window corresponds to different time granularity for example we are interested in every ten minutes for the hour before that. Each transaction in this window has weight.

- **Memory Management**

This is major issue in mining stream data. This includes choosing of efficient and compact data structure algorithm which can 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.

- **One Pass Algorithm**

There are many algorithms for mining stream data. Based on result, they are categorizing as exact algorithms or approximate algorithms. In exact algorithms, The result consist of all the itemsets which satisfy support values greater than or equal to threshold support. To Produced accurate result in stream data, additional cost is needed. In approximate algorithms, the result is approximate result with or without an error guarantee.

- **Concept Drift**

Since stream data are rapid and time varying, we cannot assume that total number of class are fixed because itemsets which are frequent can change as well with arrival of new data. So there is need of frequent updating of model, because old data are inconsistent with the new data. This problem is known as Concept drift. If we neglect non frequent itemsets from consideration which can be frequent itemset later, we cannot get this information. Therefore technique is needed to handle concept drifting.

- **Producing and Maintain Association Rules**

Mining Association rule involves a lot of memory and CPU costs. There is also one problem; processing time is limited to only one online scan. So there is need of real time maintaining and updating association rule. However stream data, if we update association rules too frequently, the cost of computation will increase drastically.

- **Resource Aware**

Resources such as memory space, CPU, and sometimes energy are very precious in data stream mining. One cannot ignore the resources availability, for example when main memory is totally used up in processing algorithm, data will be lost and it lead to inaccuracy of results. In general, if we don't consider this problem, it will degrade the performance of the mining algorithm.

APPLICATION DEPENDENT ISSUE

Based on need of application environments, association rule mining algorithms has different needs[11].

- **Timeline Query**

In some application, recent data are important while in other certain period of time data is important base on user interest. So it leads to issue of efficiently storing and retrieving with timeline.

- **Multidimensional Stream Data**

There are some applications where stream data are multi dimensional such as sensor data network. So there is need of multi-dimensional processing techniques for mining association rule. How to efficiently store, update and retrieve the multidimensional information to mine association rules in multidimensional data streams is an issue.

- **Distributed Environment**

In a distributed environment, stream data comes from multiple remote sources. So in this type of environment, there are various issues that need to be consider such as communication overhead, computation overhead, and resource overhead. Therefore there is need of distributed algorithm which imposes low communication overhead, controlled interactive response time and which can parallel and incrementally generate frequent itemsets.

- **Online Interactive Processing**

There are various algorithms which allow user to modify the mining parameters during the processing period. Therefore, how to make the online processing interactive according to user inputs before and during the processing period is another important issue.

- **Visualization**

In some data stream applications, especially monitoring applications, there is a demand for visualization of association rules to facilitate the analysis process. Visualization of data in form of graph and plot helps user to understand the relationship between related associations rules better so that they can further select and explore a specific set of rules from the visualization.

V. ANALYSIS - FPM OVER DATA STREAMS

Paper 1: Efficient Frequent Pattern Mining over Data Streams

They uses prefix tree structure called as Compact pattern Stream (CPS)-tree [20]. They use dynamic tree restructuring technique to handle stream data. For Restructuring, they use BSM method and Path adjustment method. It finds exact set of recent frequent patterns with the use of Sliding window. For each new item arrive in window, it restructure the tree. This is main disadvantage because it needs more memory and time for reconstruction.

Paper 2: Mining Frequent Itemsets in Data Streams Using the Weighted Sliding Window Model

In the year 2009, Pauray S.M Tsai proposed weighted sliding window (WSW) [21] technique. This model allow user to specify various parameter for mining like size of window, weight of window and number of window. In each window, every transaction has weight and if the weight satisfies minimum weighted threshold value then it is consider as frequent itemset. For large window size, execution time of this model decreases. This is happened because for the small window size, number of frequent itemset is small due to small number of transactions. This model uses Apriori algorithm for candidate generation and this may take more memory and time. Therefore instead of Apriori, we can use another algorithm like ‘eclate’ to improve mining.

Paper 3: Mining Frequent Itemsets over Data Streams Using Efficient Window Sliding Techniques

In the year 2009, Hue-Fu Li and Suh-Li proposed MFL-Trans SW algorithm[22] (Mining Frequent Item sets with in a Transaction Sensitive Sliding window) which is based on bit sequence and it worked on 3 phases. They are 1) Window Initialization 2) Window Sliding and 3)Pattern Generation. Based on the MFITransSW they proposed another algorithm called MFI Time SW to find the set of frequent item sets over time sensitive sliding window. Both of the technique takes more memory usage when window size increased. So hybrid approach or new technique can save time.

Paper 4: Mining Frequent Itemsets with Normalized Weight in Continuous Data Streams

In the year 2010, Yo Unghee Kim, Won young Kim and Ungmo Kim proposed an efficient technique known as WSFI mine[23] (Weighted Support Frequent Item sets mining) with normalized weight over data stream. This algorithm works in 3 phases. First phase is to divide stream data into 3 categories. They are (1) frequent items (2) latent items and (3) infrequent items. Second phase is to store compress information of frequent itemset. In the Last phase, algorithm discovers frequent itemsets. This is very efficient algorithm. This algorithm can be improve by careful pruning of infrequent itemset.

Paper 5: Mining Frequent Patterns across Multiple Data Streams

In the year 2011, Jing Guo, Peng Zhang, jianlong Tan and Li Guo proposed new algorithm called as Hybrid streaming, H-stream for short. They uses H tree for storing and maintaining historical and potential frequent itemset. It is used for collaborative and comparative frequent pattern mining [24]. Author uses real time news paper data for analyzing. Main advantage of this algorithm is that it can efficiently mine frequent pattern from multiple streams. Data may have confidential information, so one can apply privacy preserving technique to protect confidential data. It is often challenge to perform privacy preserving in high speed data stream.

Paper 6: Compression and Privacy Preservation of Data Streams Using Moments

In the year 2011, Anushree Gowtham Ringe, Deeksha Sood, Durga Toshniwal mainly focused on data compression and data privacy. They uses Moment algorithm and it works on fixed sized sliding window model[25]. This algorithm can be improved by adding noise for security purpose.

Paper 7: Mining Concept-Drifting Data Streams Using Ensemble Classifiers

Since stream data are very fast, dynamic and time varying in nature and one cannot assume that there is only fixed number of class. This is major problem because model built on old data becomes inconsistent with the arrival of new data and therefore frequent updating of the model is necessary. This problem is known as concept drifting[26]. This problem address by the author and proposed general framework which worked on drifting. The proposed algorithm combines weighted multiple classifiers by their expected prediction accuracy to mine stream data.

Paper 8: A Simple Algorithm for Finding Frequent Elements in Streams and Bags

In this paper, author uses Landmark model and stores most frequent items and their counts. It gives accurate result but for the accurate result additional cost is needed. This algorithm takes 2 scan to generate exact result items set[27]. Main advantage of this algorithm is that its memory usage is low. This algorithm discards infrequent items and their information. Infrequent items may become frequent in future therefore it needs to be stored. It requires 2 passes to generate exact result so one can improve this by single scan. In additional it gives no guarantees regarding false positive.

Paper 9: Lossy Counting Algorithm

In this paper, author proposed classic algorithm for mining known as Lossy Counting Algorithm[28]. This algorithm is based on Apriori property[2] which says that *"All nonempty subsets of a frequent itemset must also be frequent."*. It uses landmark window for processing data stream. Algorithm have good average space complexity. It generates all frequent itemset and there is no false negative.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, several issues are discussed which are needed into consideration when dealing with stream data. Due to dynamic and irregular nature of data stream, it is typical to handle stream data. Present techniques produce approximate results due to limited memory and they are not as user friendly as needed which can auto adjust some parameters like in support, confidence and error rate. As we discussed there is tread off between accuracy and response time, one should choose suitable model and algorithm based on need of application and users interest. In addition to this, we have analyzed the different existing research works of frequent pattern mining over data streams. Merits and demerits and future enhancements of the existing works are also discussed.

We can conclude that most of the current mining approaches adopt an incremental & one pass algorithm which is suitable to mind data streams, but few of them address the concept drifting [19]. The problem of handling streams is still a challenge and has wide chance of exploration for data mining researcher to carry their work. In nowadays more high-speed data streams are generated in different application domains, like millions of transactions generated from retail chains, millions of calls from telecommunication companies, millions of ATM and credit card operations processed by large banks, and millions of hits logged by popular Web sites. As most of these problems are solved and more efficient and user friendly mining techniques are developed for end users, it is quite likely that in near future data stream association rule mining play key role in business world.

REFERENCES

1. "Data mining techniques" by Arun k Pujari.
2. Aggarwal, C. (2007). In C. Aggarwal (Ed.), *"Data streams: Models and algorithms"*. Springer.
3. "Data Mining: Introductory and Advanced Topics" Margaret H. Dunham
4. Bai-En Shie a, Philip S. Yu b, Vincent S. Tseng "Efficient algorithms for mining maximal high utility itemsets from data streams ith different models" Expert Systems with Applications 39 (2012) 12947–12960
5. "Chowdary Farha ahmed, Byeong-Soo Jeong" Efficient mining of high utility patterns over data streams with a sliding window model, Springerlink.com, 2011.
6. www.borgelt.net/slides/fpm.pdf
7. "Data Streams: An Overview and Scientific Applications" Charu C. Aggarwal
8. Tanbeer, S. K., Ahmed, C. F., Jeong, B.-S., and Lee, Y.-K. 2008. "CPtree: a tree structure for single-pass frequent pattern mining's". In Proc. Of PAKDD, Lect Notes Artif Int, 1022-1027.
9. Koh, J.-L., and Shieh, S.-F. 2004. "An efficient approach for maintaining association rules based on adjusting FP-tree structures". In Lee Y-J, Li J, Whang K-Y, Lee D (eds) Proc. of DASFAA 2004. Springer-Verlag, Berlin Heidelberg New York, 417–424
10. Wang Jiinlong, Xu Conglfu, Cben Weidong, Pan Yunhe, "Survey of the Study on Frequent Pattern Mining in Data Streams", IEEE International Conference on Systems, Man and Cybernetics 2004

11. Nan Jiang and Le Gruenwald, "Research Issues in Data Stream Association Rule Mining", SIGMOD Record, Vol. 35, No. 1, Mar. 2006
12. K Jothimani, Dr Antony Selvadoss Thanamani, "An Overview of Mining Frequent Itemsets Over Data Streams Using Sliding Window Model", *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*
13. U. Chandrasekhar, Sandeep Kumar. K, Yakkala Uma Mahesh, "A Survey of latest Algorithms for Frequent Itemset Mining in Data Stream", *International Journal of Advanced Computer Research Volume-3 Number-1 Issue-9 March-2013*
14. Neha Gupta and Indrjeet Rajput, "Stream Data Mining: A Survey", international Journal of Engineering Research and Applications, Vol. 3, Issue 1, January -February 2013, pp.1113-1118
15. M.S.B. PhridviRaja, C.V. GuruRaob, "Data mining – past, present and future – a typical survey on data streams" 2212-0173 © 2013 The Authors.
16. James Cheng Yiping Ke Wilfred Ng, "A Survey on Algorithms for Mining Frequent Itemsets over Data Streams"
17. Anja Bachmann, "Challenges on Association Rule Mining On Data Streams in Contrast to Classical Association Rule Mining Algorithms" Ausarbeitungen zum Seminar, Computational Intelligence Methods Winter semester 2011/2012.
18. John Forrest, Michael Hahsler, Matthew Bolanos, "Introduction to stream: A Framework for Data", Southern Methodist University
19. Ruoming Jin, "Mining Data Streams", Ohio State University, Columbus, OH 43210
20. Syed Khairuzzaman Tabeer, Chowdary Farha ahmed, Byeong-Soo Jeong, Young Koo Lee "Efficient frequent pattern mining over data streams" 2008
21. Pauray S.M Tsai. "Mining frequent itemsets in data streams using the weighted sliding window model", Expert Systems with Applications, Vol. 36, pp. 11617–11625, Elsevier 2009.
22. Hue-Fu Li, Suh-Li "Mining frequent item sets over data streams using efficient window sliding technique", Elsevier publication. 2009.
23. Yo unghee Kim, Won Young Kim and Ungmo Kim "Mining frequent item sets with normalized weight in continuous data streams". Journal of information processing systems. 2010.
24. Jing Guo, Peng Zhang, Jianlong Tan, and Li Guo. Mining frequent patterns across multiple data streams. In Craig Macdonald, Iadh Ounis, and Ian Ruthven (Eds.), *CIKM*, 2325-2328, ACM, 2011.
25. Anushree Goutam Ringne, Deeksha Sood, and Durga Toshniwal, "Compression and Privacy Preservation of Data Streams using Moments," *International Journal of Machine Learning and Computing* vol. 1, no. 5, pp.473-478, 2011.
26. Haixun Wang, Wei Fan, Philip S. Yu, Jiawei Han; Mining Concept-Drifting Data Streams using Ensemble Classifiers; ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining; August 2003.

27. Richard M. Karp, Scott Shenker; A Simple Algorithm for Finding Frequent Elements in Streams and Bags; ACM Transactions on Database Systems; March 2003.

28. Motwani, R; Manku, G.S (2002). "Approximate frequency counts over data streams". *VLDB '02 Proceedings of the 28th international conference on Very Large Data Bases*: 346–357.

29. Chris Giannella, Jiawei Han, Jian Pei, Xifeng Yan, and Philip S Yu. Mining frequent patterns in data streams at multiple time granularities. Next generation data mining, (212):191--212, AAAI/MIT, 2003.

30. Yunyue Zhu, Dennis Shasha; StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time; Int'l Conf. on Very Large Data Bases; 2002.