A Survey on Frequent Pattern Mining Over Data Streams

S.Vijayarani[#1]
Assistant Professor,
School of Computer Science and Engineering
Bharathiar University,
Coimbatore, Tamilnadu, India
vijimohan_2000@yahoo.com

P.Sathya[#2]
M.Phil Research Scholar
School of Computer Science and Engineering
Bharathiar University,
Coimbatore, Tamilnadu, India
sathy5@ymail.com

Abstract - **Frequent pattern mining is one of the important tasks used in data mining domain. Frequent pattern mining is used to find interesting patterns from databases, such as association rules, correlations rules, sequence rules, classifier rules, and cluster rules. The main goal of the association rule is, to analyze the purchased products of a customer in a supermarket transactional data. Association rule is used to describe how frequently items are purchased together. It is mainly used in transactional data base. Data streams [12] are an ordered sequence of items that arrives in timely order. It is impossible to store the data in which item arrives. To apply data mining algorithm directly to streams instead of storing them before in a database. Real time surveillances system, telecommunication system, sensor network, financial applications, transactional data are some of the examples of the data stream systems. These types of streams produced millions or billions of updates every hour. As data stored in a database and data warehouse are processed by using some mining algorithm. Data mining [1] is defined as the process of extracting information or interesting pattern or end product from huge amount of data. In this paper, we have studied the concept of data streams and how the frequent patterns are mined from data streams. We also analyzed the existing research works in the field of frequent pattern mining data streams.**

***Keywords— Association rules mining, Data mining, Data streams, Frequent pattern mining*.**

INTRODUCTION

The main goal of the association rule is to discover all the rules that have the support and confidence greater than or equal to minimum support and confidence. When using this rule the user can omit the lower support and confidence. The association rule is used to help the retailer to develop the marketing strategies, to help to know “which items are frequently purchased by customers”. It is also used to improve the inventory management, sales management and strategy management etc.

In data streams the items are represented by record structure i.e. each individual data items may be relational tuples. Examples of some tuples are call records, web page visits, sensor reading etc. The rapid growth of continuous data has many challenges to store, computation and communication capabilities in computing system.

The high speed data needs some techniques to perform real time extraction of hidden information. Data mining is the process of finding unknown information in a large database [1] [4]. It automatically searches large stores of data to discover patterns and trends that go beyond simple analysis. It is an extraction of interesting pattern or knowledge from huge amount of data.

In data stream, data enters at a high speed rate. The system will not capable for storing the entire stream data, so it stores only a small amount of data. Data mining techniques help to find interesting patterns from unusual form of data.

Data mining techniques plays a vital role in many large organizations. But nowadays, many new techniques and algorithms are used for data streams without dropping the events. Data stream algorithms are designed with clear focus on the development of the essential data.

This paper will focus on the following sections. In Section 2, we present the various applications of data streams. Section 3 gives the overview of frequent pattern mining. Section 4 discusses the analysis of frequent pattern mining papers. Conclusion and future work of this paper are discussed in section 5.

II.DATA STREAM APPLICATIONS

Data streams are used in various applications. Some important applications are as follows [15] [1],

1. Network monitoring in data stream
2. Intrusion detection in data stream
3. Sensor network analysis in data stream
4. Cosmological application in data stream
5. Environmental and weather data in stream

a. Network monitoring in data stream:

Tele communication companies have massive streams of data containing information about phone calls. . It is important to analyze the underlying data in order to determine the broad patterns in the data. This can be extremely difficult if the number of source-destination combinations is very large.

Sketches can be used in order to determine important patterns such frequent call patterns, moments or even joins across multiple data streams.

Sketches are extremely efficient because they use an additive approach to summarize the underlying data stream.

b. Intrusion Detection in data stream:

In many applications the intrusions appear as sudden bursts of patterns in even greater streams of attacks. Intrusion makes the problem very difficult, because one cannot scan the data twice. Stream clustering turns outs to be quite useful for such problems. When known intrusions are received in the stream, they can be used in order to create class-specific clusters. These class specific clusters can be used to determine the nature of new clusters which arise from unknown intrusion detection.

c. Sensor Network Analysis in data stream:

Sensors have played an important role for collecting a variety of scientific data from the environment. So it is considered as data stream. The challenging in processing sensor data is as follows [15].

1. Sensor data may be certain in nature.
2. Data becomes very large because of different sources.
3. To extract the underlying information in sensor stream is very difficult.

Synopsis construction techniques are a natural approach for sensor problems because of the nature of the underlying aggregation

d. Cosmological Application in data stream:

The installation of large space stations, space telescope and observation result in large streams of data on different stars and clusters of galaxies. This is useful information about behavior of different cosmological objects.

The amount of data received in a single day in such application can often exceed several tera bytes. In such cases, we are using synopsis technique for compressing the data. This concept is mainly focused on the accuracy of the underlying data.

e. Environmental and Weather Data in stream:

Many satellite and scientific instruments collect environmental data such as cloud cover, wind speeds, humidity data and ocean currents. Such data can be used to make predictions about long and short term weather and climate changes. The challenge is to be able to combine these parameters in order to make timely and accurate predictions about weather driven events.

III. FREQUENT PATTERN MINING

A collection of one or more items in a transaction is known as item set. Consider the example T= {beer, bread, chips, diaper} is an item set. An item set whose threshold value is greater than equal to minimum support and confidence is known as frequent item set.

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

.

Table 1 Transaction Database

In the above table there are five transactions occurred namely A,B,C,D,E,F. In that A occurred in 3 times. B occurred in 3 times. C occurred in 2 times and D, E, F occurred 3, 4, 2 times respectively. In that example we set 3 as threshold value. So according to the threshold value A,B,D,F are called as frequent items because their occurrence values are greater than the threshold value. C and F are called as infrequent item sets. So they are omitted. Thus this is called as frequent pattern mining.

Frequent pattern mining is used in variety of areas. Some of them are [10],

1. Click stream analysis
2. Drug design
3. Market basket analysis
4. Web link analysis
5. Genome analysis

IV. ANALYSIS - FPM OVER DATA STREAMS

In the year **2008**, Syed Khairuzzaman Tanbeer, Choudary Farhan Ahmad, Byeong-Soo Jeong, Young-Koo Lee had proposed a research paper [4] "Efficient frequent pattern mining over data streams". In this paper they proposed a prefix-tree structure called CPS-tree (Compact Pattern Stream tree). The CPS tree uses a new technique called as dynamic tree restructuring technique to handle the stream data. This tree constructs a compact-prefix tree structure with single pass scanning. Its performance is as same as FP tree growth technique. After creating the CPS [13] tree we can refresh the tree at each window. For restructuring the CPS tree they used an efficient restructuring mechanism called as BSM method [14] and path adjusting method. Once the CPS tree is constructed current window the algorithm uses bottom up technique to generate exact set of recent frequent patterns.

Issues and Challenges:

In this paper they proposed prefix-tree structure called CPS-tree that introduces dynamic tree restructuring mechanism in data stream and find recent frequent patterns. The main disadvantage of this algorithm is every time a new item is arrives, it reconstructs the tree. So it causes more memory space as well as time. In future we should discover some new reconstructing techniques to avoid these problems.

In the year **2009** Pauray S.M. Tsai proposed research paper [5] “Mining frequent item sets in data streams using the weighted sliding window model”. In this paper the author proposed a new technique called the weighted sliding window WSW algorithm. This model allows the user to specify the number of windows for mining, the size of the window and the weight each window. Using this algorithm the user can specify minimum weighted threshold value. They split the transaction into equal number of windows. Using the WSW algorithm they calculate the weight of each transaction in each window. If the weighted support count of an item is greater than or equal to minimum weighted threshold value it is called as frequent item set. Using the Apriori algorithm the user can generate the candidate item set also. When a candidate item set is generated we can determine whether it is frequent or not by using the WSW algorithm.

Issues and Challenges:

The main advantage of this algorithm is, it scanned the database only once. It does not take more than one scan to find out the frequent item set. When the window size increases, then the execution time of WSW decreases. This is because when the window size small the number of transaction containing frequent item sets in each window is small. Therefore the probability of choosing a candidate item set to be not a frequent item set of early windows is high. The candidate item set generation may take more time and memory. In this paper they used Apriori algorithm for candidate generation. In future we may use D-éclat or rapid mining algorithm instead of Apriori for avoiding this candidate generation.

In the same year **(2009)** Hue-Fu Li and Suh-Li had proposed a research paper called [6] “Mining frequent item sets over data streams using efficient window sliding techniques”. In this paper they proposed an efficient bit-sequence based algorithm called MFI-Trans SW (Mining Frequent Item sets with in a Transaction Sensitive Sliding window). MFI algorithm worked on three phases. They are Window Initialization, Window Sliding and Pattern Generation. In Window Initialization phase, every item of each transaction is encoded in an efficient bit sequence representation. In second phase the algorithm uses the left bit shift sequence technique to slide the windows efficiently. In final phase the complete set of frequent item sets within the current sliding window is generated. Based on the MFI-TransSW they proposed another algorithm called MFI-TimeSW to find the set of frequent item sets over time sensitive sliding window.

Issues and Challenges:

This algorithm scanned the database only once to find the frequent item sets. The proposed MFI-TransSW algorithm is a time efficient method for mining frequent item sets from data stream with in a transactional sensitive sliding window. If the window size increased, the memory usage of MFI-TransSW is also increased. As the same, the window size increases, the processing time of phase 1 and phase 2 of MFI-TimeSW is also increased. To construct these two algorithms may take more time. In future we may combine these two into one algorithm for saving time.

4. Due to the characteristics of data stream there are lots of problems arise for mining stream data. In the year **2010**, Yo Unghee Kim, Won young Kim and Ungmo Kim had proposed a research paper called [7] “Mining frequent item sets with normalized weight in continuous data streams”. They proposed an efficient algorithm WSFI mine (Weighted Support Frequent Item sets mining) with normalized weight over data stream.

The proposed WSFI-mine method is designed to mine all frequent item sets from one scan in the data stream. This algorithm uses three phases. In the first phase the data stream is divided into 3 categories such as frequent items, latent items, and infrequent items. In second phase the author present a novel tree structure, called WSFP-tree that stores compressed crucial information about frequent item sets. WSFP-tree structure is an extended form of FP-tree growth technique. At last phase the WSFI-mine method discovers frequent item sets.

Issues and challenges:

This WSFI-mine algorithm can mine all frequent item sets in one scan from the database. Data changes usually with time in the data stream. A currently infrequent pattern may become frequent in the future. Therefore the author has to be careful not to prune infrequent item sets too early. This is the main advantage of this algorithm. There are lots of confusion arises after the pruning. In future we should discover some new techniques to avoid this confusion at the time of pruning. In future we may set some threshold values in the starting stage, to correctly divide the frequent, infrequent, and latent item sets.

In the year **2010** Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer and Byeong-Soo Jeong had proposed a research paper called [8] "Efficient mining of high utility patterns over data stream with a sliding window model". In this paper they proposed a novel algorithm for sliding window based high utility pattern mining over data stream called as HUPMS (High Utility Pattern Mining in Stream data).

This algorithm used a novel tree structure called HUS tree. It maintained the streaming data in the form of sorted order and maintains batch-by-batch information. This tree used a property called "build once mine many" i.e. building the data structure only once, several mining operation can be done. This tree also uses the concept of pattern growth mining approach, for building level wise candidate generation. Using the "build once mine many" property, every time a new item arrive the user can rebuild the tree by changing the threshold value.

Issues and challenges:

The main contribution of this paper is to provide a novel algorithm for sliding window based high utility pattern mining over data stream. This HUMPS algorithm can capture only the recent change of knowledge in a data stream by using novel tree structure. It is easy to construct and maintain the tree during the sliding window-based stream data mining. It uses the property called build once mine many. This is only suitable for interactive mining. This paper also proposed based upon FP-tree growth. So in future we may propose a new technique instead of FP –tree approach.

6. In the year **2011** Jing Guo,Peng Zhang,jianlong Tan and Li Guo discussed how to mine frequent patterns across multiple data streams called [9] "Mining frequent patterns across multiple data streams". In this paper they selected real time news paper data for analyzing. In multiple streams it is important to discover collaborative frequent patterns and comparative frequent patterns. Collaborative frequent pattern means to report frequent item in all media. This paper news section is considered. Comparative frequent pattern means to report more frequently appeared items in a media than others. It is impossible to find these two problems under one solution. To overcome these problems they proposed new algorithms called hybrid steaming, H-stream for short. This algorithm built a new hybrid frequent tree to maintain historical frequent and potential frequent item sets for efficient collaborative and comparative pattern mining. H-tree worked on 3 steps.

First step: A-tree maintains frequent and potential item sets discovered from all streams. In second step: A-hybrid window which records the frequencies of the maintained item sets in each time interval. In third step: A header table using an array of pointers the frequent item sets can be effectively retrieved from tree.

Issues and challenges:

There are lots of frequent item set mining algorithms have been proposed. But they are mined only single stream data. In many real world applications stream data are generated from multiple sources. So it is necessary to combine multiple data stream for mining. This algorithm is proposed to discover comparative and collaborative frequent patterns across multiple data streams. Using this algorithm the user can be effectively retrieved the frequent items. Using this one algorithm the user can easily retrieve the comparative and collaborative frequent item sets. Data streams data are from various sources, and it has much confidential information also so we can protect these confidential data by applying a privacy technique in future. It is often a challenge to perform privacy for continuously arriving data.

In the year **2011** Anushree Gowtham Ringe, Deeksha Sood, Durga Toshniwal had proposed a research paper [11] "Compression and privacy preservation of data streams using moments". It was mainly focused on how to prevent the misuse of sensitive data, in a stream. In this paper they proposed a novel technique for preserving the privacy of data stream, so along with providing an additional incentive of data compression.

This paper mainly focused on data compression as well as privacy. It is mainly based on data compression and privacy preservation using the concept of moment. In this concept, they used fixed size window keeps sliding with the arrival of new data pointers in the stream. In each window they assigned a moment of the curve generated by its data points. The set of centroids can then be used for subsequent analysis of data streams. The data points are then replaced by a single point called as moment. Also it is not possible to retrieve the original data points from the moment hence providing privacy.

Issues and challenges:

In this paper a new technique has been proposed which provides data compression as well as privacy preservation in data streams. In this paper they used only one real time example gold price for mining. To calculate the moment value is very difficult because of the large volume of data. In future we can extend this technique to Boolean data as well. This technique can also be extended to multi- variant data streams. In addition for security purpose, we can add noise for further encryption.

V.CONCLUSION AND FUTURE WORK

Data streams data and sensor data are also becoming richer. 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. Mining techniques will then be very significant in order to conduct advanced analysis, such as determining trends and finding interesting patterns, on streaming data.

In this paper we have studied the concept of data streams and how the frequent patterns are mined over data streams. 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. In future, we will develop new techniques and algorithms for finding frequent patterns over data streams which helps to overcome the drawbacks of the existing techniques.

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)”Efficient frequent pattern mining over data streams” Syed Khairuzzaman Tabeer, Chowdary Farha ahmed, Byeong-Soo Jeong, Young Koo Lee 2008

5) “Mining frequent item sets in data streams using the weighted sliding window model”Pauray S.M.Tsai, 2009, Elsevier publication

6)“Mining frequent item sets over data streams using efficient window sliding technique” Hue-Fu Li, Suh-Li, 2009, Elsevier publication.

7) “Mining frequent item sets with normalized weight in continuous data streams” Yo-unghee Kim, Won Young Kim and Ungmo Kim 2010. Journal of information processing systems.

8*)* “Efficient mining of high utility patterns over data streams with a sliding window model” Chowdary Farha ahmed, Byeong-Soo Jeong, 2011. Springerlink.com

9) “Mining frequent patterns across multiple data streams” Jing Guo, Peng Zhang, Jianlong Tan and li Guo, 2011

10) www.borgelt.net/slides/fpm.pdf

11)”Compression and privacy preservation of data streams using moments”Anushree Gowtham Ringe, Deeksha Sood and Turga Toshniwal, 2011. Information journal of machine learning and computing.

12) “Data Streams: An Overview and Scientific Applications*”* Charu C. Aggarwal

13) Tanbeer, S. K., Ahmed, C. F., Jeong, B.-S., and Lee, Y.-K. 2008. “CP-tree: a tree structure for single-pass frequent pattern mining”*S*. In Proc. of PAKDD, Lect Notes Artif Int, 1022-1027.

14) 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

15)” Data Stream Mining A Practical Approach”Albert Bifet, Geoff Holmes, Richard Kirkby and Bernhard Pfahringer May 2011

16) “Scientific Data mining and Discovery”, ISBN 978-3-642-02787-1. Springer Verlag Berlin Heidelberg, 2010.

Mrs. S.Vijayarani has completed MCA and M.Phil in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy and security issues and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Ms. P.Sathya has completed M.Sc in Software Systems. She is currently pursuing her M.Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are Data Streams in data mining and privacy preserving in Data mining.