A Survey on Mining Frequent Itemsets over Data Streams

Shailvi Maurya
Assistant Professor
School of Information Technology, ADYPU
Pune, Maharashtra (India)

Sneha Ambhore
Assistant Professor
School of Information Technology, ADYPU
Pune, Maharashtra (India)

Sneha Parit
Assistant Professor
School of Information Technology, ADYPU
Pune, Maharashtra (India)

ABSTRACT
Mining frequent itemsets over data stream has been challenging task. The incoming data from various sources like ecommerce website, click streams, text, audio, weather forecasting etc. are massive unbounded and high speed that it is impractical to store all, process and scan complete data at the same time to extract information. While processing memory and time are the main parameters must be minimum consumed. Thus the paper provides different algorithms for mining over static and dynamic data also known as data stream.

Keywords
Data mining, data stream, frequent itemsets.

1. INTRODUCTION
Data mining [1] is considered as a process of discovering useful patterns beneath the data, also uses machine learning algorithms. There have been techniques which used computer programs to automatically extract models representing patterns from data and then check those models. Traditionally used data mining techniques are not applied to data streams because most required multiple scans of data to extract information, for stream data it was unrealistic. Information systems have been more complex, even amount of data being processed has increased and are dynamic in nature due to continues updates. In data mining the main focus is on cleansing of data which is known as noise elimination or noise reduction.

As the technology is getting advanced so the data generated is unbroken unbounded tremendous and high speed called as data stream which often changes with infinite period. Data streams mining is the most demanding problems in data mining. Real time applications produce large amount of data streams, such as sensor data generated from sensor networks, online transaction flows in retail chains, Web record and click-streams in Web applications, call records in telecommunications, and performance measurement in network monitoring and traffic management[6]. Following data mining requirement must be satisfied:

- **Memory utilization:** Data stream is generated continuously and massive so the memory usage should be low for low cost.
- **Fast computing:** The computing speed should be high for better performance because slow computing will lead to high time consumption.
- **Data Scan:** The scanning of data must be done once due to the dynamic nature of the data stream and limited amount of memory.
- **Updating Data Stream:** The updated result of scan and analysis should be available fast on user's appropriate request.

Data streams are dynamic in nature which changes with time. The changing nature of stream is known as concept change. Concept change is the change of the underlying concept over time. The concept change occurs when the data generating process changes from one data generating model to another one. Concept change can be categorized in two main categories, concept drift and concept shift. Concept drift describes a gradual change of the concept and concept shift happens when a change between two concepts is more abrupt [9].

The discovery of the sets of items that frequently appear in a transaction database refers to mining frequent patterns. Mining frequent patterns (or itemsets) [2] has become common and fundamental problem in the context of unearthing knowledge and data mining. Because of wide application area in business, industry and science. The well-known algorithm for this known situation is the Apriori algorithm [1] for static databases. Frequent itemset mining in static databases is a well-established problem in data mining community for which many fast and scalable algorithms have been proposed. An itemset which is also known as set of items is frequent in a database if the number of its occurrences in the database is more than or equal to a user specified threshold. This threshold is specified by the user of the mining process [9]. The fig.1 shows frequent mining of items.

To mine data stream efficiently multiple scans are not acceptable. Mining on recent data is more attractive than the old history in data streams. It is challenging to mine frequent itemsets from recent data, because new items coming and old items overdue with high speed [7].

The frequent pattern mining is not limited to static databases but it is extended to dynamic databases and data streams [11]. Mining frequent itemsets from a transaction database is a fundamental task to several data mining applications. Mining frequent itemsets is one of the most important problems in data stream research area. Mining in data streams incurs extra challenges.

2. RELATED WORK

In this section various algorithms for mining frequent itemsets over static or dynamic data was developed were Apriori algorithm [1] based on association rule was proposed by Agrawal and Srikant for static database. The algorithm discovered association rule between large set of dataset of sales transactions. The proposed hybrid algorithm called as AprioriHybrid. The algorithm has excellent scale-up properties with respect to transaction size and items in database.

Mohammed J. Zaki proposed algorithm for association rule known as Eclat [2] algorithm (Equivalence CLAss Transformation) for static database which uses efficient traversing techniques to identify the long frequent itemsets quickly which utilized the structural properties of frequent itemsets for fast discovery. The items are organized into lattice which further divides into sub-lattice which can be solved in memory.

The estDec [3] method based on data stream which is dynamic in nature was proposed by Chang et. al.. The algorithm finds recent frequent itemsets over an online data stream by decaying the weight of old transactions as time goes by. As a result, the recently change of information in a data stream adaptively reflected the current mining result of the data stream. The weight of information in a transaction of a data stream is slowly reduced as time goes by while its reduction rate can be flexibly controlled. Therefore, no transaction needs to be maintained physically..

Another algorithm on data stream was proposed by Leung et.al. for extracting useful information. The algorithm proposed a novel tree structure, called DSTree (Data Stream Tree) [4], which hold important data from the stream. It uses window concept where tree captures the contents of transaction in a window for processing and arranges nodes to canonical order which is unaffected by changes in item frequency. The DSTree can be easily retain and mined for frequent patterns and different other patterns like constrained itemsets

Due to real-time response and computational complexity Mozafari et. al. which introduced a verification algorithm SWIM (Sliding Window Incremental Miner) [5] based on sliding window for mining over data stream. This algorithm improves performance by allowing small report delay. Thus the proposed algorithm shows greater efficiency, flexibility and scalability for mining frequent itemsets on data stream with larger windows. Therefore this algorithm is fast verifier to simplify the association-rule mining problem under the realistic assumption

Demanding properties, like unknown or limitless size, possibly fast arrival rate, inability to trace back previously arrived transactions, and lack of system control over the order in which the data arrive Hua-Fu Li et. al. proposed an effective bit-sequence based, one-pass algorithm, called MFI-Trans-SW (Mining Frequent Itemsets within a Transaction-sensitive Sliding Window) [6], to mine itemsets frequency from data streams in a transaction-sensitive sliding window which holds a fixed number of transactions. The proposed algorithm depicts accuracy, faster processing and less memory consumption.

Another algorithm based on prefix tree based structure called as BFI- tree [7] was proposed by Kun Li et. al. to maintain accuracy in frequent itemsets from sliding windows over stream of data. Tracking the limits between frequent itemsets and infrequent itemsets, it restricts update on small part of the tree. It is time effective and returns accuracy in results by outperforming MFI-TransSW in time and space.

The paper proposed efficient technique Compact Pattern Stream tree (CPS-tree) [8] developed by Tanbeer et. al. discovered complete set of recent pattern which are frequent due to high speed data stream over sliding window. For finding frequent patterns from recant data enhances the analysis of data stream It also introduced the concept of dynamically tree restructuring in CPS-tree to produce frequency- descending tree structure at runtime. The algorithm is efficient in time complexity and low memory consumption.

Sliding window model is widely used for mining over data stream for mining recent data. Mahmood Deypir et.al. proposed algorithm VSW (variable Sliding Window) [9] where the window size is dynamically adjusted based on the amount of the concept change arriving within the data stream. The window expands as the concept becomes stable and shrinks when a change in concept detected. When the algorithm effectively detects concept change, adjust the window size and adapts dynamically to new concept.

Fatemeh Nori et. al. proposed the concept of Tmoment [10] algorithm. The algorithm uses data structure for storing transactions of the window and corresponding frequent closed itemsets. The Tmoment algorithm mines closed frequent itemsets within sliding window over data stream. It updates the new transactions arrived and old transactions are departed. This algorithm is suitable for high-speed and unbounded transactional data streams.

Popular association rule mining algorithms such as Apriori algorithm which generate a huge number of candidate 2-itemsets. To remove these drawbacks, a new algorithm named Matrix Based Algorithm with Tags (MBAT) [12] is proposed by Harpreet Singh and Renu Dhir which finds the frequent itemsets directly from the transactional matrix which is generated from the database to generate association rules. Proposed algorithm greatly decreases the number of candidate itemsets, mainly candidate 2-itemsets.

Apriori based techniques, Frequent Pattern growth (FP-growth) and Equivalence CLASS Transformation (ECLAT) are the approaches used mostly in extracting frequent patterns. These mining methods consume more calculation time and very difficult to implement. Fast frequent pattern mining technique using vertical stream in the form of binary representation algorithm VFFM [13] which is considered to be more easily executable and also effectively implemented.

Table 1. Comparison analysis of frequent itemsets mining algorithms

Name of Algorithm	Processing Time (sec)	Memory Utilization (MB)	Implementation Tool	Data scan	Application area
Apriori	0.25	30	C	Multiple Scan	Static Database
Prefix based	30	1.5	Sun Ultra-2 Work Station	Multiple Scan	Static Database
estDec	0.25	15	C	Multiple Scan	Online Database
DSTree	60	30	C++	Multiple Scan	Real Time
SWIM	0.2	15	C	Single Scan	Streaming
MFI-Trans SW	50	15	Microsoft visual C++ 6.0	Single Scan	Real Time
BFI-stream	30	30	C++	Single Scan	Online Database
CPS-tree	600	10	Microsoft visual C++ 6.0	Single Scan	Online Database
VSW	0.01	0.60	C++	Single Scan	Real Time
Tmoment	0.001	0.70	C++	Single Scan	Real Time
MBAT	0.01	20	C	Single Scan	Static Database
VFFM	0.4	1.5	Java 1.2.0	Single Scan	Static Database

3. CONCLUSION

The aim of the study was to survey different algorithm used for mining over data stream providing in depth knowledge of data stream. As due to high volume incoming data generated from various applications, data stream its becoming typical to handle because of unbounded, massive, high-speed and variant nature. Mining.

The survey paper is comparison of various mythologies for mining frequent itemsets over static and dynamic database. The paper gives guidance on the different frequent mining algorithms based on different parameter which will be helpful in further research work semantic annotation for frequent patterns, and contextual analysis of frequent patterns.

4. ACKNOWLEDGMENTS

A special thanks to Sneha Ambhore and Sneha Parit for supporting and helping in survey and making the comparative analysis of different mining algorithms.

5. REFERENCES

[1] Agrawal, R., & Srikant, R. 1994. “Fast algorithms for mining association rules”, In Proc. VLDB int. conf. very large databases (pp. 487–499).

[2] Zaki, M. 2000. “Scalable algorithms for association mining”, IEEE Transactions on Knowledge and Data Engineering, 12(3), 372–390.

[3] Chang, J., & Lee, W. S. 2003. “Finding recently frequent itemsets adaptively over online transactional data streams”, Information Systems, 31(8), 849–869.

[4] [Leung, C. K.- S., & Khan, Q. I. 2006. “DSTree: A tree structure for the mining of frequent sets from data streams”, In Proc. ICDM (pp. 928–932).

[5] Mozafari, B., Thakkar, H., & Zaniolo, C. 2008. “Verifying and mining frequent patterns from large windows over data streams”, In Proc. int. conf. ICDE (pp. 179– 188).

[6] Li, H.-F., & Lee, S.-Y. 2009. “Mining frequent itemsets over data streams using efficient window sliding techniques”, Expert Systems with Applications, 36(2), 1466–1477.

[7] Li, K., Wang, y. y., Ellahi, M., Wang, H.-an 2008. “Mining recent frequent itemsets in data streams, IEEE fifth Int. conf. on Fuzzy Syatem and Knowledge Discovery”.

[8] Tanbeer, S. K., Ahmed, C. F., Jeong, B. S., Lee, Y. K. 2009. “Sliding window-based frequent pattern mining over data streams”, Information Sciences 179 (2009) 3843–3865

[9] Deypir, M., Sadreddini, M. H., Hashemi, S. 2012. “Towards a variable size sliding window model for frequent itemset mining over data streams”, Computers & Industrial Engineering 63 (2012) 161–172.

[10] Nori, F., Deypir, M., Sadreddini, M. H., Hashemi, S. 2013. "A sliding window based algorithm for frequent closed itemset mining over data streams", The Journal of Systems and Software 86 (2013) 615– 623.

[11] Deypir, M., Sadreddini, M. H., Hashemi, S. 2011. "A dynamic layout of sliding window for frequent itemset mining over data streams", The Journal of Systems and Software 85 (2012) 746– 759.

[12] Harpreet Singh, Renu Dhir 2013. "A New Efficient Matrix Based Frequent Itemset Mining Algorithm with Tags", International Journal of Future Computer and Communication, Vol. 2, No. 4.

[13] C.Ganesh, B.Sathiyabhama, T.Geetha, 2016. "Fast Frequent Pattern Mining Using Vertical Data Format for Knowledge Discovery", International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-5, Issue-5).