Advances in Modelling and Analysis D
Vol. 23, No. 1, December, 2018, pp. 27-30
Journal homepage: http://iieta.org/Journals/AMA/AMA_D

A survey on mining frequent item sets from data stream

Bhargavi Peddireddy[1*], Anuradha Ch[2], Sri R. Chandra Murty Patnala[1]

[1] Department of Computer Science and Engineering, ANUCET, Acharya Nagarjuna University, Guntur 522510, A.P, India.
[2] Department of Computer Science and Engineering, V.R. Siddhartha Engineering College, Vijayawada 520007, A.P, India

Corresponding Author Email: srirampatnala@gmail.com

https://doi.org/10.18280/ama_d.230105

Received: 10 October 2018
Accepted: 5 December 2018

Keywords:
data streams,FPM, data base, itemset

ABSTRACT

Data mining is a process of finding undisclosed, unknown and interested patterns from databases. It has led to the many techniques, and they emphasis on mining frequent patterns to capture patterns whose occurrence is frequent, unusual and rare occurrence patterns. The research has led to optimize the performance of the techniques with its applications. But, traditional data mining techniques are limited to the databases which exhibits static behavior. But, the real time applications like sensors and stock data exhibits the behavior where the incoming data speed is fast and the cumulated data is huge. Such kind of databases are named as Data streams. The compatibility of data streams with the applications has led to the many issues and challenges. It has been motivated researchers to propose various frameworks and algorithms to speed up the mining process. In this paper, we discuss the various Applications of Data Streams, issues, and challenges. We discuss various models such are Landmark, Sliding window, Damped, and Title timed widow models. And also discuss various frequent itemset mining Algorithms for each models. In addition, this paper also discusses research issues and future direction towards for variety of pattern mining.

1. INTRODUCTION

Frequent Pattern Mining (FPM) is a process of analyzing transactions of a data base for deriving hidden knowledge which is a pair of items that are occurred together in more transactions. FPM plays an important role in Association rule mining [30], hence it led to several techniques [1-3, 13]. FPM techniques are classified into two categories that are Candidate generate and test based approaches and Frequent Pattern Growth (FPG) approaches [14]. Applications made researchers to extend FPM to further, some are efficient in-terms of space, time and passes, compact representations [2] [5], ordered patterns [12], sequential patterns [11-12, 14]. And also it is extended [13] to the data bases that are exhibiting dynamic nature, where few transactions are added and few transactions are deleted. It is also interested to use FPM techniques in another variant of data base is data stream.

Data stream is defined as a concept of continuous, high-speed generated data that challenges computer to store, analyze to take strategic decision in real time. Due to the continuous flow of data, the discovered patterns may be violated for incoming coming data. And the rare patterns whose occurrence is frequent at some portion may be violated. Means that the traditional data mining techniques of static databases are failed to extract knowledge from data streams. It has motivated researchers to apply frequent pattern mining concepts and new techniques for aiming the applications of data streams.

The purpose of this paper is to give survey on the applications, challenges of data streams and issues, various techniques developed for mining frequent patterns from data streams, analysis and comparative study of those techniques.

This Paper is organized as follows: preliminaries, applications and issues of data streams is presented in section 2. Frequent pattern mining strategies are presented in section 3, Research issues and future directions are described in the section 4 and section 5. It is concluded with conclusion section.

2. PRELIMINARIES

Let I ={i_1, i_2, … , i_n} be a set items used so far in the domain DB, T be a transaction that contains a set of itemsets and T_k be the k^{th} transaction of a database DB={T_1, T_2, … ,T_k}. An itemset ***e*** be a set of items, where eϵ(2^I-∅), |DB_k| denotes the total number of transactions in the database DB. δ be the minimum support threshold given by the user. Tm be a new transaction generated at this instance and it is represented as T_m= {i_1, i_2, … , i_n}.this it shows DB allows addition of new transactions.

SUP(e) is defined as the occurrence of an itemset in transactions of DB, where the occurrence once is counted in each transaction.

An itemset Fe is said be a frequent, if the occurrence of F_e or *SUP(F_e)* in the database is greater than or equal to the given threshold value. The problem is to mine such patterns whose occurrence is greater than the given threshold value.

2.1 Applications

Network: Telephone Network allows users to communicate with each other. For better security, it is necessary to Analysis such kind of data to identify the hidden underlying patterns. Hence companies need to record the information about phone calls as data streams.

Intrusion Detection: Technology made users to active in many applications that leads to huge storage and causes to abrupt eruption of patterns. Data must be checked rigorously to detect intrusion that reads data repeatedly. But in data streaming, the data cannot be scanned twice which makes the problem more cumbersome.

Sensor Network Analysis: Sensors are used to collect a large amount of data from the nature or environment that can be considered as data stream. Usually, Sensor data is quite ambiguous in nature which makes difficult in the underlying information. Hence efficient techniques are required to store, processing and analyzing.

Environmental and Weather Data: Usually, sensors are used to collect vast amounts of data from environment that is weather data as data stream to predict climate changes. The challenging task is to find the parameters which can help to predict possible changes. Hence efficient techniques are required to predict the weather data.

2.2 Issues

i. The primary issues is to have the extraction model to extract the data from the source and extract knowledge as patterns from source. The proposed extraction model should match the speed of the data model which is fast, continuous and unbounded. Hence there are three models available Landmark, damped window and sliding window model.

ii. The incoming data are high speed and continuous. Generally FPM approaches relatively slower than data streams. Hence buffering management techniques helps to solve this issue, where incoming data will be in queue when mining process busy with previous data. Hence an efficient methodologies are required to handle the arrival data.

iii. Data streams comes from more than one source with different speed. The computation overhead on handling such data is high. Hence intelligent techniques which can extract and mine data with exhibiting parallel and incremental techniques.

iv. Data streams exhibits dynamic nature, means that incoming data comes like flooding data. Hence the extracted knowledge will be updated transaction by transaction. It is required more storage space and mechanism to count the patterns. The extracted knowledge may become invalid over the incoming data. Hence efficient techniques are required are required to handle itemsets storage.

v. Due to the continuous flow of data, the mined frequent pattern need to be updated for every new data. Efficient data structures are required to handle such kind of data.

3. MINING FREQUENT PATTERNS FROM DATA STREAMS

In this section, we discuss the various models and algorithms for FPM from the data mining research. FPM has been attracted to mine frequent patterns from data streams. To exhibit continuous flow of the data, several extensions have been proposed to FPM to extend it to the data streams. But all approaches had challenges.

I. The current frequent items sets may become infrequent in future.

II. The current itemsets which are not interested may become interested patterns or frequent in future.

In order to mine item sets from data streams, new concept is introduced named as window model, whose size was fixed or variable with the transactions from n^{th} to m^{th}. It is observed from the literature that, various models have been proposed those are presented below.

3.1 Land mark window model

In this model, the approach divides the stream into series of buckets where each bucket consider the transactions from *i* to *n* for mining procedure that gives all approximated frequent patterns from *W(i,n)*. Here, item set, frequency, and the itemset upper bound as error are maintained. And the proposed algorithm uses Buffer to keep incoming transactions that can be used to find the frequency of an itemset and remove itemsets whose support does not reach threshold, *Trie* is a forest to keep the above information of itemsets maintained in each level of pre-order traversal of forest *trie*. *Trie* gets updated for each new transaction and old one gets deleted, And *set-gen()* function is to generate all possible candidate itemsets. This approach requires a lot of space and computation time.

Yu et al. [11] proposed *Chernoff bound* [9] based false negative approach which does not allow false positive which is deriving frequent patterns. It is efficient than the above trie based approach, since it requires less space and computation.

Lee et al. [12] has proposed a single pass approach with ISFI tree structure to derive frequent patterns from data streams. It maintains Header table and sub frequent Item sets trees are constructed for every item in the transaction.

3.2 Sliding window model

The motivation behind the use of sliding window is to derive frequent patterns from the recent transactions of data streams are named as recent frequent patterns. This idea has been used in many applications thus lead to the various approaches [11-12]. Sliding window process only the items in a window that length decided by the application and maintains only frequent information. It slides window transaction by transaction or a set of transactions. It is difficult to use the above error rate and false positive result of the previous window, because it returns a large number of false positive item sets.

Chi et al. [8] has proposed heuristic based approach to generate condensed lossless represented patterns named the closed frequent itemsets. Their approach maintains closed patterns of the sliding window in a *CET* (Closed Enumeration Tree) called Moment. It is biased to the size of main memory which is based on assumption that the transaction can be accommodated in main memory. Their structure helped in maintaining substantial for successive sliding windows in data streams, but not enough to meet high speed incoming data.

Tanbeer et al. [2] proposed two phase algorithm. In first phase, items of sliding window are stored in a Compact tree structure *CPS* Tree as predefined order. In the second phase, the tree is restructured to derive all closed frequent itemsets. It is efficient than *CET* approach, but it is to be reconstructed for each new item which takes more space and more computation time.

Tai et al. [13] proposed weight based sliding window approach to mine frequent patterns from data streams. In this approach, weight is assigned to each window, number of windows are assigned to the data as per the importance and minimum support per each window. Single phase approach

WSW is proposed to derive frequent patterns. But it shows high impact when window size gets decreased only.

3.3 Damped window model

One of the notable approach is *estDes* proposed by Chang and Lee [15], it is aimed at overcoming the issue of old transaction domination with the consideration of decay rate rather than ignoring support of item sets while window is moving in forward direction. Basically it is an approximated approach with four phases. In first phase, parameters are updated, counts of itemsets are updated in second phase, itemsets are updated with respect to the delay rate in third phase, and finally frequent patterns are determined. Delay factor reduces the old transaction domination over the newly found itemsets. However, it can produce a large error rate while applying the apriori property.

3.4 Titled time window model

The motivation behind for this approach is to mine frequent patterns from continuous forward natured sliding window approach. The basic idea of this approach is to maintain frequent patterns at various granularities in *FP-stream* data structure. FP-stream maintains two components, prefix tree to maintain patterns using FP tree and tilted window.

Han et al. [4] proposed an approach consists of two trees, the first one is *FP tree* maintains patterns of the tilted window and the second one is pattern tree which maintains frequent information of the past windows. The applications of data streams shows that mining patterns from data streams is a demanding issue. Many applications need to find the variation occurring among the patterns which are highly time sensitive. Landmark model may not be suitable efficient to handle such patterns. Some applications may be interested in analyzing recent data, literature proves that sliding window based methods suitable for extracting such patterns. However it is biased to the main memory, means that they are efficient as long as main memory fits the data. It has led to the many techniques single pass algorithms.

4. RESEARCH ISSUES

1. Data streams are a sequence of ordered transactions. The challenging issue is to extract variation of frequent item sets over the transactions.
2. Data streams are unbounded data, it is challenging to use memory efficiently for maintaining large amount of frequent information, data structures to compute the cumulative support.
3. Incoming data rate is high, it is challenging issue to match the execution speed of the proposed approaches. Hence it is required to have an approach which takes less number of passes over the data stream.

5. FUTURE DIRECTIONS

I. **Data structure**: Memory issue can be resolved with the use of efficient data structure, which can store item sets efficiently such that computation and space gets reduced. Development of such kind of data structures are necessary to improve the efficiency of the proposed approaches.

II. **Incremental Mining**: Since data streams exhibits continuous and dynamic nature, dynamic and incremental data structures are necessary to maintain patterns and association rules which can be changed stream by stream.

III. **Multiple resources**: Since data streams comes from many sources, it demands efficient approaches to handle when data is coming from multiple resources to avoid join which is costly method.

IV. **Multidimensional Mining**: The traditional mining methods are efficient in analyzing one dimensional data, whereas the real time data streams may exhibit multidimensional data. Hence, this demands efficient approaches [14] to handle such kind of data.

V. **Visualization of rules and patterns**: some applications may demand the visual representation of frequent and association rules for better understanding. Hence efficient techniques are required to handle such kind of approach.

VI. **Tuple Evaluation**: In data streams, the traditional approaches works on assumption that is the streams are new transactions. But in real time, streams may be updated version of old stream. It is challenging issue for the above models to revisit the old stream.

VII. **Rare patterns** [6-8, 10]: Rare patterns could give a variety of patterns whose occurrence in some transactions. Efficient techniques are required to handle such patterns.

VIII. **Outlier detection** [5]: Outlier detection plays an important role in improving the efficiency by avoiding unnecessary transaction in data analyzing. Efficient methods are required to handle such kind of data

6. CONCLUSIONS

Itemset mining and Data stream mining are the active fields in research having scope in various applications. This paper addressed the basics of FPM, and Data streams environment. Also presented various applications with issues and challenges in it. This paper has discussed various models for handling data steams issues, and algorithms for mining frequent itemsets. At end, paper has presented research issues, and future directions for variety and other forms of knowledge.

REFERENCES

[1] Rakesh A, Srikant RK. (1995). Mining sequential patterns. Data Engineering. Proceedings of the Eleventh International Conference on IEEE, pp. 3-14. http://dx.doi.org/10.1109/ICDE.1995.380415

[2] Saha B, Lazarescu M, Venkatesh S. (2007). Infrequent item mining in multiple data streams. Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), pp. 569-574. https://doi.org/10.1109/ICDMW.2007.32

[3] Giannella C, Han J, Pei J, Yan X, Yu PS. (2003). Mining frequent patterns in data streams at multiple time granularities. Next Generation Data Mining 212: 191-212.

[4] Hemalatha CS, Vaidehi V, Lakshmi R. (2015) Minimal infrequent pattern based approach for mining outliers in data streams. Expert Systems with Applications 42(4): 1998--2012. https://doi.org/10.1016/j.eswa.2014.09.053

[5] Huang D, Koh YS, Dobbie G. (2012). Rare pattern mining on data streams. International Conference on Data Warehousing and Knowledge Discovery, LNCS 7448: 303-314. https://doi.org/10.1007/978-3-642-32584-7_25/

[6] Huang DTJ, Koh YS, Dobbie G, Pears R. (2014). Detecting changes in rare patterns from data streams. PAKDD-2014, LNAI 8444, pp. 437-448. http://dx.doi.org/10.1007/978-3-319-06605-9_36

[7] Manku GS, Motwani R. (2002). Approximate frequency counts over data streams. Proceedings of the 28th International Conference on Very Large Data Bases, pp. 346-357. http://dx.doi.org/10.1016/B978-155860869-6/50038-X

[8] Karnati R, Subramanyam RBV. (2016). Efficiently maintaining and discovering sequential patterns with sequence deletion using discovered sequences. International Journal of Applied Engineering Research 11(1): 685-691. https://www.ripublication.com/ijaer16/ijaerv11n1_104.pdf

[9] Deypir M, Sadreddini MH, Hashemi S. (2012). Towards a variable size sliding window model for frequent itemset mining over data streams. Computers & Industrial Engineering 63(1): 161-172. https://doi.org/10.1016/j.cie.2012.02.008

[10] Jian P, Han JW, Mao RY. (2000). CLOSET: An efficient algorithm for mining frequent closed itemset. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery 4(2). https://doi.org/10.1109/CSSE.2008.1042

[11] Jin R, Agrawal G. (2007). Frequent pattern mining in data streams. Data Streams: Models and Algorithms. http://dx.doi.org/10.1007/978-0-387-47534-9_4

[12] Tanbeer SK, Ahmed CF, Jeong BS, Lee YK. (2009). Sliding window-based frequent pattern mining over data streams. Information sciences 179(2): 3843--3865. http://dx.doi.org/10.1016/j.ins.2009.07.012

[13] Lee VE, Jin R, Agrawal G. (2014). Frequent pattern mining in data streams. Frequent Pattern Mining, Springer, pp. 199-224. http://dx.doi.org/10.1007/978-3-319-07821-2_9

[14] Chi Y, Wang H, Yu PS, Muntz RR. (2004). Moment: Maintaining closed frequent itemsets over a stream sliding window. Data Mining, ICDM'04. http://dx.doi.org/10.1007/s10115-006-0003-0

[15] Kim YH, Kim WY, Kim UM. (2010). Mining frequent itemsets with normalized weight in continuous data streams. Journal of Information Processing Systems 6(1): 79-90. http://dx.doi.org/10.3745/JIPS.2010.6.1.079