Volume 5, Issue 9, September 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Frequent Pattern Mining From Fast Changing Complex Data Streams

Rajiv Senapati, D. Anil Kumar
Department of CSE, GIET, Gunupur,
Odisha, India

Abstract— Data Stream Mining is one of the most important areas which is growing high in today's business world. Stream Mining is a process of extracting knowledge from continuous data records. Now a day's many sources are there which generates huge number of data rapidly in a continuous fashion, which includes: sensors, telephones, mobile devices, satellites, GPS system, internet, online marketing, telemedicine system, stock market, weather forecasting, entertainment sector, banking sector, credit cards, social media like face book, twitter, LinkedIn, etc. This paper addresses the issues of frequent pattern mining in static dataset and its limitation as well as frequent pattern mining in dynamic datasets, which we call as data stream mining.

Keywords— Static Dataset, Dynamic Dataset, Apriori, Stream Mining, Sliding Window

I. INTRODUCTION

Data Mining is a process of extracting hidden pattern from huge amount of data, where the data is basically static in nature. It will analyse the historical data and extract desirable knowledge. But, now a day the data are changing dynamically and growing rapidly. That is the reason why some extra effort has to be given to track the incoming data and analyse properly to produce interesting useful patterns. Some of the well known fundamental frequent pattern mining algorithms are already proposed [6] based on BFS and then FP-Growth [7] based on DFS. Based on these basic fundamental algorithms many other algorithms are being proposed in later time. In this paper we are trying to differentiate the frequent pattern mining in static datasets and dynamic data sets which is termed as data streams.

Static Datasets: We have studied simple and structured data sets such as data in relational data bases, transactional data bases and data warehouses. Those data sources can be considered as structured or semi structured some times and all are static in nature, where data mining operation can be performed easily for pattern evaluation. Some of the existing algorithms like Apriori [6], FP-Growth [7] is sufficient to find out the frequent pattern and associations among data.

Dynamic Datasets: These data sets are basically in complex form for example: semi-structured, unstructured, spatial and temporal, hypertext and multimedia etc. Some of the sources of these kinds of datasets are:

- ✓ Satellite –mounted remote sensor that is constantly generating data.
- ✓ Telecommunication data, where data related to all call made, call time, source, destination, call duration, call type, message etc are constantly generating.
- ✓ Transaction data from retail industry.
- ✓ Data from electric power grids.
- ✓ Time-series data relating to stock market.
- ✓ Sales forecasting.
- ✓ Utility studies.
- ✓ Observation of natural phenomena such as atmosphere, temperature, wind etc.
- ✓ Banking and Credit card data.
- ✓ Social media like Face book, Twitter, Link data. etc.
- ✓ World Wide Web
- ✓ Bioinformatics
- ✓ Real time surveillance system
- ✓ Internet Traffic
- ✓ Industry production process etc.

All these data are massive e.g. terabytes in volume, temporally ordered, fast changing, and infinite. These data are dynamic in nature and also referred as stream data. Traditional data mining methods are not capable to analyse dynamic stream datasets. Because it requires multiple scans of the data and therefore not applicable for stream data analysis. Unlike static datasets, stream data flows in and out of a computer system continuously. So it is impossible to store the data and scan it multiple times. In this paper a new data structure is discussed called sliding window model, which is very useful where only recent events may be important. It also reduces memory requirements because only a small window of data is stored for analysis.

Frequent pattern mining has been utilized in extensive applications such as medical and bio data analysis [1], stock market and protein networks [2], network environment [3], traffic data analysis [4], analysis of web click stream [5] and so on.

II. RESEARCH ISSUES IN MINING FREQUENT PATTERNS FROM STATIC DATASETS AND DYNAMIC DATASETS

Data is the key concern in today's business world and data analysis plays a major role in making business process smooth. However data sets which we are considering for analysis are of two kinds, these are: Static Type and Dynamic Type. Static datasets sometimes referred as non stream dataset are historical in nature, where data mining algorithms can be applied for pattern evaluation and future forecasting, whereas dynamic datasets focuses more on the current scenario. Some of the research issues related to static datasets are:

i. Data Pre-processing: Data pre-processing is always being an important criteria to be taken care in data analysis task because the real world data is highly susceptible to noisy, inconsistent and missing values. Low quality data leads to low quality result that is the reason why data has to be processed and converted into a proper format so that it can be considered for analysis. Dimensionality reduction is also a main focus because all the data coming from different sources many not useful for our analysis.
ii. Memory constraint: The requirement of memory space for storing data is not a big problem now days, but when analysis comes into picture then it is very difficult to analyse the data from huge amount of data records. The reason is: static data mining requires scanning of the entire datasets multiple times for analysis. Where as in data stream mining only one time scan is sufficient to evaluate the frequent patterns. This technique is useful when the current data plays vital role in analysis. Here choosing the appropriate data structure is one of the major concerns.

III. FREQUENT PATTERNS FROM DYNAMIC DATASETS

As data have been accumulated more quickly in recent years, corresponding databases have also become huge. Hence general frequent pattern mining methods have been faced with limitations that do not appropriately respond to the massive data. To overcome this problem, few data mining methods have been studied which conduct more efficient mining tasks by scanning databases only once. The sliding window approach is proposes which perform mining operation by focusing on recently accumulated parts over data streams [8].

Frequent pattern mining can be applied in both static and dynamic data streams. Data stream means that the transactional data are added continuously. The data stream mining has to satisfy the following requirements [9].

i. Each data element needed for data stream analysis has to be examined only once.
ii. Because the data elements are continuously added, the memory uses for mining operations should be limited to an acceptable range.
iii. All the entered data has to be processed as soon as possible.
iv. The results of data stream analysis should be available instantly as well as their quality should also be acceptable when ever users need the result.

Algorithm – 1

Sliding window-based frequent pattern mining over data streams

The FP-Growth approach is efficient for static databases, but it is not suitable for data streams with continuous data flow. The FP-Growth method scans the dataset more than two times hence they do not deal with data streams instantly. The algorithm constructs tree with items remained after infrequent items are deleted, they have to discard previously generated trees and build new trees again if new transaction data are added in the data stream. The two scan based method must read databases from the first again since they already eliminated infrequent items in the previous step. To solve this mining method suitable for data streams [10, 11] have been proposed and they can perform mining task with only one database scan, thereby responding to changes of data stream immediately.

After that, sliding window-based frequent pattern mining approaches [11, 12, 13, 14, 15, 16] have been proposed, which can mine frequent patterns considering the latest transaction data of large data streams. Especially in those paper an efficient tree-restructuring method, BSM was proposed. The method performs restructuring operations more effectively than previous ones such as the path adjusting method, etc. IWFP algorithm [12] a weighted frequent pattern mining algorithm over data streams, applying the BSM method. Among accumulated data streams, the most important elements are recently added data in general. In other words, importance of previously added data can be lowered or meaningless, while that of lately accumulated ones can be relatively higher. Therefore, to reflect these characteristics, the sliding window model can be applied into mining process. The method divides data streams into windows composed of a set of constant-sized transactions and finds frequent patterns from recently generated windows, where the size of windows and the number of them can be assigned as various values by users. Through the sliding window-based approach, we can always obtain frequent patterns reflecting recent information. In [17], Tanbeer et al. suggested a frequent pattern mining algorithm over sliding window-based data streams, applying the BSM technique to tree restructuring steps in order to raise efficiency of mining operations.

Algorithm – 2

Maximal frequent pattern mining over data streams

Mining all frequent patterns over data streams may leads to numerous computational overheads in general if data sizes are huge. In sliding window-based data stream mining, since the remaining parts except for the latest windows are not considered, the overheads can be reduced, but we cannot still avoid causing them if the size of windows or the number of

them becomes large. In MAFIA algorithm (Burdick et al., 2005), vertical bitmap representation was proposed so as to help mine MFPs more efficiently. The algorithm uses an additional data structure with a bitmap form to reduce the number of tree traversals. After the bitmap is constructed, MAFIA can know pattern's frequency through AND operation of the bitmap even though it does not try to traverse trees actually. FP max (Grahne & Zhu, 2005) is a state-of-the-art MFP mining algorithm, where FP-array, an additional data structure for mining MFPs more quickly, was proposed, thereby decreasing tree traversal times considerably. Since FP-array has information of patterns' supports, the algorithm can calculate them in advance before trees are actually traversed when growth processes are performed. Consequently, this technique not only can reduce tree traversal operations effectively but also can enhance pruning efficiency by preventing generation of needless conditional trees. However, since the above algorithms have two scan-based processes, they are not suitable for the data stream mining.

Algorithm – 3

Applying weight conditions into frequent pattern mining over data streams

Each item existing in data streams has certain weight. For instance, given items over retail data streams, support information of them mean their sales volume, and their weight information represents prices or profits for each item. Therefore, when both of those two elements are considered, we can gain mining results reflecting complex factors in the real world. Weights of items in data streams are used in the mining process after they are converted into normalized values within a certain range. The reason is that if a weight of any item is too large, it is hard to denote its weighted support as a finite number of digits. The main challenge of applying weights is to maintain the anti-monotone property. However, the application generally destroys that property since weighted infrequent patterns can become weighted frequent ones as pattern growth operations are conducted. For this reason, researchers have made efforts to maintain the anti-monotone property, and a variety of methods. WFPMDS (Ahmed et al., 2009) mines weighted frequent patterns over data stream environment based on the sliding window model. The algorithm conducts tree restructuring work with the BSM technique and provides the most recent mining results from the sliding window whenever users request them. In this study, the framework of the proposed algorithm, WMFP-SW is based on the state-of-the-art MFP mining algorithm, FPmax and the outstanding tree restructuring technique, BSM.

Algorithm - 4

Weighted maximal frequent pattern mining over data streams based on sliding window model

Consider a data stream shown in table 1 and represented in Fig.1

Table 1. Data stream with weight information of the data

TID	Transaction	Item	Weight
100	I2, I5	I1	0.5
200	I3, I5	I2	0.7
300	I1,I2, I3,I4,I7	I3	0.8
400	I2,I3,I6	I4	1.0
500	I1, I2, I3, I4, I5, I6	I5	0.4
600	I2, I5, I6	I6	0.9
700	I1, I3, I4, I5	I7	0.6
800	I1, I4, I5	I8	0.3
900	I2, I3, I4, I8		

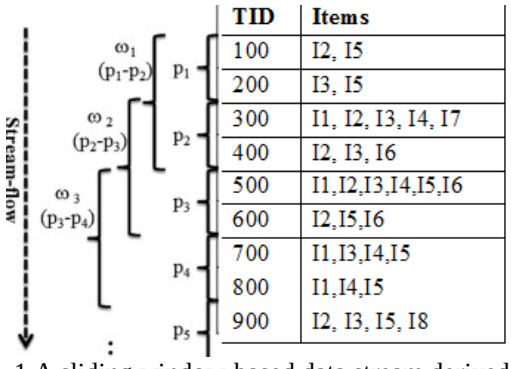

Fig. 1.A sliding window-based data stream derived from Table 1

In the above scenario both the window size (i.e. the number of panes) and pane size (i.e. the number of items) are set as 2. In the data stream, sliding window process is performed as follows. We first read $p1$ and $p2$ to fill $\omega 1$. After that, when reading the next pane, $p3$, we remove the old pane, $p1$. Then, $p2$ and $p3$ belong to the current window, $\omega 2$. In the same manner, the old pane, $p2$ is deleted and the new pane $p4$ is entered into the current window, $\omega 3$ in the next step. As a

result, the current window always has the latest data stream information, and thus, the sliding window-based mining methods can instantly provide users with frequent pattern results considering the most recent data whenever they request mining results.

IV. CONCLUSIONS

In this paper we have investigated various data mining algorithm for static dataset mining as well as for data stream mining. Here we have studied an algorithm for weighted maximal frequent pattern over data stream based on the sliding window model. The studied algorithms can be applied in various areas like closed frequent pattern mining, high utility pattern mining etc.

REFERENCES

[1] Sallaberry, Arnaud, et al. "Sequential patterns mining and gene sequence visualization to discover novelty from microarray data." Journal of Biomedical Informatics 44.5 (2011): 760-774.

[2] Sim, Kelvin, et al. "Mining maximal quasi-bicliques: Novel algorithm and applications in the stock market and protein networks." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 2.4 (2009): 255-273. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.

[3] Fang, Guodong, Zhihong Deng, and Hao Ma. "Network traffic monitoring based on mining frequent patterns." Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on. Vol. 7. IEEE, 2009.

[4] Liu, Wei, et al. "Discovering spatio-temporal causal interactions in traffic data streams." Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011.

[5] Li, Hua-Fu. "A sliding window method for finding top-k path traversal patterns over streaming web click-sequences." Expert Systems with Applications 36.3 (2009): 4382-4386.

[6] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.

[7] Han, Jiawei, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation." ACM SIGMOD Record. Vol. 29. No. 2. ACM, 2000.

[8] Lee, Gangin, Unil Yun, and Keun Ho Ryu. "Sliding window based weighted maximal frequent pattern mining over data streams." *Expert Systems with Applications* 41.2 (2014): 694-708.

[9] Farzanyar, Zahra, Mohammadreza Kangavari, and Nick Cercone. "Max-FISM: Mining (recently) maximal frequent itemsets over data streams using the sliding window model." Computers & Mathematics with Applications 64.6 (2012): 1706-1718.

[10] Ahmed, Chowdhury Farhan, et al. "An efficient algorithm for sliding window-based weighted frequent pattern mining over data streams." IEICE TRANSACTIONS on Information and Systems 92.7 (2009): 1369-1381.

[11] Ahmed, Chowdhury Farhan, et al. "Single-pass incremental and interactive mining for weighted frequent patterns." Expert Systems with Applications 39.9 (2012): 7976-7994.

[12] Chen, Hui, et al. "Mining frequent patterns in a varying-size sliding window of online transactional data streams." Information Sciences 215 (2012): 15-36.

[13] Deypir, Mahmood, Mohammad Hadi Sadreddini, and Sattar Hashemi. "Towards a variable size sliding window model for frequent itemset mining over data streams." Computers & Industrial Engineering 63.1 (2012): 161-172.

[14] Farzanyar, Zahra, Mohammadreza Kangavari, and Nick Cercone. "Max-FISM: Mining (recently) maximal frequent itemsets over data streams using the sliding window model." Computers & Mathematics with Applications 64.6 (2012): 1706-1718.

[15] Zhang, X., and Y. Zhang. "Sliding-window top-k pattern mining on uncertain streams." Journal of Computational Information Systems 7.3 (2011): 984-992.

[16] Tanbeer, Syed Khairuzzaman, et al. "Efficient single-pass frequent pattern mining using a prefix-tree." Information Sciences 179.5 (2009): 559-583.

