International Journal of Soft Computing 3 (2): 144-146, 2008
ISSN: 1816-9503

Efficient Graph Structure for the Mining of Frequent Itemsets From Data Streams

E.R. Naganthan and F. Ramesh Dhanaseelan
Department of Computer Science and Engineering, Alagappa University, Karaikudi

Abstract: In this study, we propose a graph structure, that captures important data streams. This graph can be easily maintained and mined for frequent item sets as well as various other patterns like constrained item sets. This graph captures the contents of transaction in a window and arranges nodes according to some canonical order that is unaffected by changes in item frequency. This graph structure is designed for exact stream mining of regular frequent item sets.

Key words: Stream mining, data stream, association rule mining, frequent itemsets

INTRODUCTION

A data stream is an ordered sequence of items that arrives in timely order. Different from data in traditional static databases, data streams are continuous, unbounded, usually come with high speed and have a data distribution that often changes with time (Guha *et al*., 2001). One example application of data stream association rule mining is to estimate missing data in sensor networks (Klatcher and Gruenwald, 2005). Mining from data the following 2 properties of streams is more challenging due to data streams, ie., the data streams are continuous and unbounded and data in the streams are not necessarily uniformly distributed; their distributions are usually changing with time. In recent years, several stream mining algorithms have been proposed and they can be broadly categorized into 2 classes, i.e., exact and appropriate algorithms. Exact algorithms (Chi *et al*., 2004) find truly frequent item sets and appropriate algorithms (Giannella *et al*., 2004) find frequent item sets by using appropriate procedures, i.e., these algorithms may find some infrequent item sets or may miss some frequent item sets. We propose a graph structure, which is designed for exact stream mining of regular frequent item sets. The graph captures the contents of relevant transactions in the streams. When the streams flow through, a fixed size user window containing the interesting portion of the streams, i.e., the recent data is properly updated.

Traditional association rule mining algorithms are developed to work on static data and, thus, can not be applied directly to mine association rule in stream data (Gaber *et al*., 2005; Jiang and Gruenwald, 2006; Chang and Lee, 2004). The first recognized frequent item sets mining algorithm for traditional databases is Apriori (Agarwal *et al*., 1993). After that, many other algorithms based on the ideas of Apriori were developed for performance improvement (Hans *et al*., 1999). Apriori-based algorithms require multiple scans of the original database, which leads to high CPU and I/O costs.

Therefore, they are not suitable for a data stream environment, in which data can be scanned only once. Another category of association rule mining algorithms for traditional databases proposed by Hans *et al*. (2000) are those using a Frequent Pattern tree (FP-tree) data structure and an FP-growth algorithm which allows mining of frequent item sets without generating candidate item sets. Compared with Apriori-based algorithms, it achieves higher performance by avoiding iterative candidate generations. However, it still can not be used to mine association rule in data streams (Leung and Khan, 2006; Charikar *et al*., 2004) since the construction of FP-tree requires 2 scans of data.

GRAPH STRUCTURE FOR DATA STREAM

The graph structure is designed for exact stream mining (Yu *et al*., 2004; Gulab and Ozsu, 2003). The construction of the graph structure only requires on scan of the streaming data. The graph structure captures the contents of transactions in each batch of streaming data.

We arrange transaction items according to some canonical order, which can be specified by the user prior to the graph construction or the mining process. For example, items can be consistently arranged in lexicographic order or alphabetical order. Alternatively, items can be arranged according to some specific order depending on the item properties-such as their price

Corresponding Author: E.R Naganthan, Department of Computer Science and Engineering, Alagappa University, Karaikudi

values or their validity to some constraints, which can also be determined prior to the graph construction or the mining process. We keep a list of frequency counts at each node.

Whenever a new batch of transactions flows in, we append to this list at each node its frequency count in the current batch. In other words, the last entry of the list at node X shows the frequency count of X in the current batch. When the next batch of transactions comes in, the list is shifted forward. The last entry shifts and becomes the second-last entry; this leaves room for the newest batch. At the same time, the frequency count corresponding to the oldest batch in the window is removed (Lin *et al*., 2005). This has the same effect as deleting from the window the transactions in the oldest batch.

We use a pointer to indicate the last update at each node. If the pointer points to the previous entry in the list of frequency counts at a node X, then this indicates that X has just been visited at the update of the last batch. On the other hand, if the pointer points to a much earlier entry in the list at a node Y, then this indicates that Y has not been visited since then and that the frequent counts of Y for the entries in-between should be 0s.

Since the graph structure is constructed independent of minsup, every transaction in the current window is captured. Once such a tree is constructed, we can mine frequent itemsets from it in a fashion similar to FP-growth (Han *et al*., 2006) (using minsup). Since, items are consistently arranged according to some canonical order, one can guarantee the inclusion of all frequent items using just upward traversals. There is also no worry about possible omission or doubly -counting of items during the mining process. Consequently, we find all and only those truly frequent itemsets because we use minsup for mining and because every transaction in the current window is captured in the graph structure.

To summarize, transaction items are arranged according to some canonical order in our graph structure so that the ordering is unaffected by the changes in frequency caused by the continuous nature of streams. When the window slides, transactions in the oldest batch can be easily detected by shifting the list of frequency counts. The effective use of pointer at each node helps us avoid performing the expensive tree traversal of all nodes. Moreover, mining is delayed until it is needed. Since our graph structure is always kept up-to-date, frequent itemsets in current streams can be found effectively. By using such a delay evaluation scheme, we avoid lots of unnecessary computation. As streams are continuously flowing rapidly, computation spent on older batches of transactions may have been wasted if these batches get removed from the current window before the user mines for frequent itemsets.

Table 1: The graph structure after each batch of transaction is added for stream mining

Batch	Transactions	Contents
First	t_1	{a,b,c}
	t_2	{a}
	t_3	{a,c}
Second	t_4	{a,c,d}
	t_5	{b,d}
	t_6	{a,b,d}
Third	t_7	{b,d}
	t_8	{ a,b,c,d}
	t_9	{a,c}

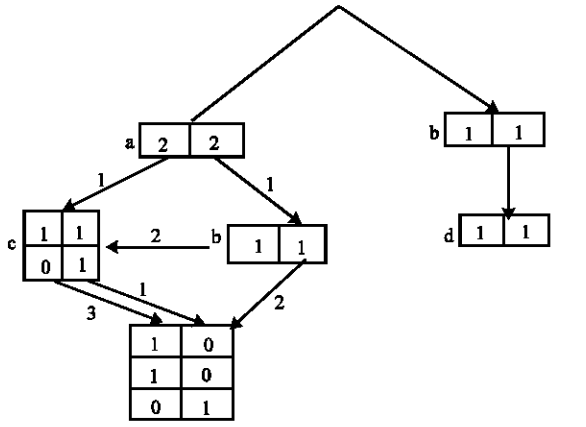

Fig. 1: At time T (the graph structure capturing 1st and 2nd batches) At time T′ (the graph structure capturing 2nd and 3rd batches)

Let minsup be 3 and let the window size *w* be 2 batches (indicating that only 2 batches of transactions are kept). Then, when the first 2 batches of transactions in the streams flows in, we insert the transactions into our graph structure and keep frequency counts in a list of *w* entries at each node. Each entry in the list corresponds to a batch. For example, the node a: 3: 2 in Table 1 indicates that the frequency of a is 3 in the first batch and is 2 in the second batch.

Afterwards (at time T′) when the third batch of streaming data flows in, we insert the transactions in our graph structure. The list of frequency counts shifts, the frequent counts for the oldest (i.e., the first) batch are removed-leaving room for the frequency counts for the second and the third (i.e., the two newest) batches of transactions. For example, if we call the mining process at time T′, we get frequent itemsets (a): 4, (a,c): 3, (a,d): 3, (b,d): 4, (c): 3 and (d): 5 (Fig. 1).

RESULTS AND DISCUSSION

In the experiments, we mainly evaluated the accuracy and efficiency of our graph structure. In the first experiment, we measured the accuracy of graph structure.

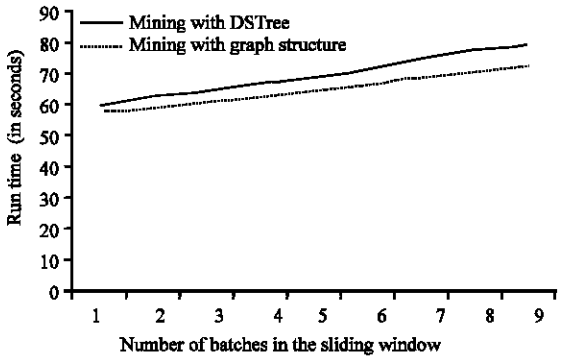

Fig. 2: Number of batches

As the graph structure captured the w most recent batches of transactions in the streams, we compared the frequent item sets returned by mining directly form these transactions with those returned by mining from our graph structure. The experimental results show that mining from our graph structure led to 100% accuracy. In other words, mining from the graph structure returned all and only those truly frequent item sets. All the returned item sets were frequent, and all frequent item sets were returned. This shows that our graph structure, which is designed for an exact stream mining. The graph structure captures the contents of transactions in the streams.

In the second experiment, we measured the efficiency of our graph structure. We compared the runtime of mining from our graph structure with that of using the DSTree. The x-axis shows the number of batches in the current window. The y-axis shows the run time. Figure 2 shows the run time. When the number of batches increased, the run time of mining from our graph structure slightly increased. But it is better than DSTree.

The result of third experiment show that, our graph structure required less space than the DSTree.

CONCLUSION

This graph structure captures the contents of transactions in a window, and arranges tree nodes according to some canonical order that is unaffected by changes in item frequency. Mining from this graph structure returned all and only those truly frequent item sets. All the returned item sets were frequent, and all frequent item sets were returned. So this graph structure is suitable for exact stream mining of regular frequent item sets.

REFERENCES

Agarwal, R., T. Imielinski and A. Swami, 1993. Mining Association Rule between Sets of Items on Massive Databases. Int'l Conference on Management of Data.

Chang, J.H. and W.S. Lee, 2004. A Sliding Window Method for Finding Recently Frequent Itemsets over online Data Streams. Journal of Information Science and Engineering.

Charikar, M., K. Chen and M. Farach-Colton, 2004. Finding frequent items in data streams. Theoretical Computer Sciences.

Chi, Y. *et al.*, 2004. Moment: Maintaining closed frequent item sets over a stream sliding window. In proc. ICDM, pp: 59-66.

Gaber, M.M., A. Zaslavsky and S. Krishnaswamy, 2005. Mining Data Streams: A Review. SIGMOD Record, Vol. 34.

Giannella, C. *et al.*, 2004. Mining frequent patterns in data streams at multiple time granularities. In Data Mining: Next generation challenges and Future Directions, AAAI/MIT Press, pp: 6.

Golab, L. and M.T. Ozsu, 2003. Issues in Data Stream Management. In SIGMOID Record, Vol. 32.

Guha, S., N. Koudas and K. Shine, 2001. Data stream and Histogram. ACM Symposium on Theory of Computing.

Han, J. *et al.*, 2000. Mining frequent patterns without candidate generation. In Proc. SIGMOID, pp: 1-12.

Hans, J., G. Dong and Y. Yin, 1999. Efficient mining of partial periodic patterns in time series database. IEEE Int'l Conference on Data Mining.

Hans, J., J. Pei and Y. Yin, 2000. Mining Frequent Patterns without Candidate Generation: Int'l Conference on Management of data.

Jiang, N. and Le Gruenwald, 2006. Research Issues in Data Stream Association Rule Mining. SIGMOID Record, Vol. 35.

Kai-Sang Leung, C. and Q.I. Khan, 2006. DSTree: A Tree Structure for the Mining of Frequent Sets from Data Streams. IEEE 6th Int'l. Conference on Data Mining.

Kalatcher, M. and Le Gruenwald, 2005. Estimating Missing values in Related Sensor Data streams. Int. Conference on Management of data.

Lin, C.H., D.Y. Chiu, Y. Wu and L.P. Arbee Chen, 2005. Mining Frequent item sets from Data Streams with a Time-Sensitive Sliding Window. SIAM Int'l Conference on Data Mining.

Yu, J.X. *et al.*, 2004. False positive or false negative: Mining frequent item sets from high speed transactional data streams. In Proc. VLDB, pp: 204-215.