Comparative Approach of Sequential Progressive Database with Respect to Static and Incremental Database

[1]Mr. Nayan S. Mahalle

Information Technology

PRMIT&R, Badnera.

[2]Prof. Mrs. S. S. Sikchi

Information Technology

PRMIT&R, Badnera.

Abstract

Data Mining is the process of automatically searching large volumes of data for patterns. It includes various technique of recognizing frequently appearing pattern of number, stream, and various databases like static, incremental and progressive pattern mining. This paper presents the review of sequential progressive pattern mining and need of it in comparison with static and incremental approach also focus on work on progressive sequential pattern mining. Sequential pattern mining is a method for determining time-related behavior in sequence databases. It is obvious timestamp as an important attribute of each dataset, it is important factor and it can give us more accurate and useful information over changing period. Although there have been many recent studies on the sequential patterns in static and incremental database. But complexity of sequential pattern mining increases as data increases on dynamic basis over changing time. Decision support problem motivates data mining which is faced by most retail organization. A sequence database consists of sequences of ordered elements or events, recorded with or without a concrete notion of time.

Keywords: Sequential pattern mining**,** Progressive mining.

1. Introduction:

Knowledge discovery in databases (KDD), or Data Mining, is the effort to understand, analyze, and eventually make use of huge volume of data available**.** Data mining is the discovery of hidden information found in databases and can be viewed as a step in overall process of Knowledge Discovery in databases (KDD) [1][2].

With the rapid growth in size and number of available databases in commercial, industrial, administrative and other applications, it is necessary examine how to extract knowledge automatically from huge amount of data for different type of analysis [3].It is the integration of various techniques from multiple disciplines such as statistics, machine learning, pattern recognition, neural networks, image processing, and database management system and so on [4].

Mining is an important subject of data mining, a further promotion of association rule mining, and it is also widely applied [5]. Sequential pattern mining algorithms address the problem of discovering the existent frequent sequences in a given database [6]. Sequential patterns can help to determine which items are bought one after another in a sequence [7], or to analyze browsing orders of homepages in a Web site [9] and more.

The sequential pattern mining on progressive databases is very new approach which progressively discovers the sequential patterns in period of interest. Period of interest is sliding window which continuously advances as time goes. New item are inserted to the dataset of interest and old which are not frequent are removed from it and become up to date. In general, the existing proposals do not fully explore the real world scenario, such as items associated with support in data stream applications such as market basket analysis.

2. Pattern Mining on a Static Database

Data Mining that uses static database for mining is known as static data mining. There are different static data mining algorithms like Apriori, Fp-Tree, Fast algorithm, Partition based algorithm etc. Static data mining algorithms like Apriori, Fp-Growth, Fast Algorithm, Partition Based Algorithms are applies on original database. If there is a need to update or delete some or all the existing set of data during the process of data mining then repetition of whole

procedure is required. This type of repetition is time-consuming becomes cause of lack of efficiency [14].

3. Pattern Mining on a incremental Database

In incremental data mining, the entry in database increases continuously as new item is inserted into it and along with old data it directly adds newly arrived item with old one. So, size of database increase and it stores as combination of old and new item or entries in dataset. While extracting such big database it represents older pattern which are not needed as data changes on the fly and user of database is always interested in new pattern rather than old one. Due to this finding sequential pattern in incremental database may lack of interest to the user [12]. Using frequent sequences and support counts discovered from original database, *FASTUP* rapidly updates frequent sequences and their counts by scanning over increment database instead of whole updated database. Fewer but more promising candidates are generated by just checking counts in increment database [11].It is noted that users are usually more interested in the recent data than the old ones. If certain sequence does not have any newly arriving elements that sequence will still stay in database and undesirably contribute to database. New sequential patterns which appear frequently in the recent sequences may not be considered as frequent sequential patterns.

4. Need for mining of progressive database

The assumption of having a static database may not hold in many applications. The data in real world usually change on the fly. When we deal with an incremental database, it is not feasible to remine the whole sequential patterns every time when the database increases because the remining process is costly. To handle the incremental database, Parthasarathy et al. presented the algorithm ISM [10] using a lattice framework to incrementally update the support of each sequential pattern in equivalent classes. Masseglia et al. derived the algorithm ISE [11] to join candidate sequential patterns in original database with the newly increasing database. Cheng et al. introduced algorithm IncSpan [12], which utilized a special data structure named sequential pattern tree to store the projection of database.

However, the incremental mining algorithms can only handle the incremental parts of the database. Because of the limitation of data structures maintained in algorithms, it can only create new candidates but cannot delete the obsolete data in a progressive database. The deletion of an item from the database results in the reconstruction of all candidate item sets, which induces incredible amount of computing.

In practice, users are usually more interested in the recent data than the old ones. To capture the dynamic nature of data addition and deletion, a model of sequential pattern mining with a progressive database while the data in the database may be static, inserted, or deleted [13].

5. Sequential Pattern mining

Sequence Pattern Mining is the mining of frequently occurring ordered events or subsequences as patterns. An example of sequential pattern is "Customers who buy a product are likely to buy other product within a month". For retail data, sequential patterns are useful for shelf placement and promotions. Also telecommunications and other businesses may also use sequential patterns for targeted marketing, customer retention and many other tasks. Other areas in which sequential patterns can be applied include Web access pattern analysis, weather prediction, production processes, and network intrusion detection analysis. Most studies of sequential pattern mining concentrate on categorical patterns. The sequential pattern mining problem was first introduced by Agrawal and Srikant in 1995[8]

6. Progressive Database

Using various techniques on progressive database it is possible to add new data, delete old data and to find various frequent data pattern on time basis. Although there have been many recent studies on the mining of sequential patterns in a static database and in a database with increasing data, these works, do not fully explore the effect of deleting old data from the sequences in the database. When sequential patterns are generated, the newly arriving patterns may not be identified as frequent sequential patterns due to the existence of old data and sequences [13].

7. Progressive Sequential Pattern Mining

The data in real world changes on the fly. Moreover, finding sequential patterns in an incremental database may cause lack of interest to the users. When sequential patterns are generated, the obsolete sequential patterns that are not frequent recently may stay in the reported results. The incremental mining algorithms do not consider the deletion of the obsolete data from the sequence database. That is, these works are not applicable to a progressive database. However, if a certain sequence does not have any newly arriving elements, this sequence will

still stay in the database and undesirably contribute to the number of sequences in the sequence database. Therefore, when new sequential patterns are generated, the new patterns which appear frequently in the recent sequences may not be considered as frequent sequential patterns because number of sequences in the sequence database is never reduced. In view of this, the infrequent sequential patterns whose timestamps are obsolete should be removed [13].

Sequential pattern mining with a progressive database is widely used in many fields. For example, prediction of prefetching data to a mobile user on the wireless gateway is an essential application [15].Whenever a mobile user asks an item from a wireless gateway, the prefetching system decides which other items the mobile user may want. By an accurate and predictive mechanism, the prefetching system can significantly improve the query latency to mobile users. Moreover, the stock price changes of the company which went into bankruptcy five years ago may have very little influence on the quotes of the other stocks now. The applications mentioned above are suitable to apply progressive sequential pattern mining techniques.

An efficient algorithm PISA, which stands for Progressive mIning of Sequential pAtterns, corresponding to the mining in a progressive database. Pisa takes the concept of period of interest (POI) into consideration.POI is a sliding window, whose length is a user specified time interval, continuously advancing as the time goes by. The sequences having elements whose timestamps fall into this period, POI, contribute to the number of sequences in the sequential database for current sequential patterns. On the other hand, the sequences having only elements with timestamps older than POI should be pruned away from the sequence database immediately and will not contribute to the sequence database thereafter [13].

8. PS-Tree

PS-Tree is the core part of the algorithm PISA. It contains the information of all sequences in a progressive database and helps Pisa to generate frequent sequential patterns in each POI. There are two types of nodes in the PS-Tree: root node and common node. Root node, as the root of PS-tree, contains a list of common nodes as its children. Each common node stores its node label (element of the sequence) and a sequence list (list of sequence IDs to represent the sequences containing this element). Each sequence ID in the sequence list is marked by a corresponding timestamp [13].

. **9. Progressive Mining Algorithm:**

Initially new elements can be added to database with sequence ID, current timestamp which may be relevant or any other sequences When new elements arrives at the timestamp, say t+1, it traverses the original PS-tree of timestamp t in post order and updates the PS-tree of timestamp t [13].

As the first step, progressive mining gets the elements of all sequences at current timestamp and traverses the PS-tree. Then, it moves forward to the next timestamp until there is no newly arriving element in a progressive database. The main procedure, traverse, is used to traverse the PS-tree of timestamp t in the post order and transform it to the new PS-tree of timestamp t+1.The basic idea of the procedure traverse is to append each newly arriving element of all sequences into PS-tree.

For each node, Progressive mining examines all sequences which have new elements. If the node is root, it checks all the combinations of all the newly arriving elements in the current time. For each candidate element, if the element already appeared at the previous timestamps, it is one child of root node. Then, it has to check if the sequence ID has already existed in that child node. If the sequence ID has been in the sequence list, it updates the timestamp of that sequence to be the new timestamp. Otherwise, progressive mining creates a new sequence and inserts the sequence into the sequence list of that child. If there is no child node with the same label, it creates a new child node with the corresponding sequence ID in the sequence list. In this way, subsequences containing only one element inserted in PS-tree.

If the node being processed is not root node, first deletes the obsolete sequences in the sequence list. If there is no sequence ID left in the sequence list, progressive mining removes this node away from its parent and goes to next node. On the other hand, if there are still some sequences left, it checks them in the sequence list instead of checking all newly arriving sequences as processing root node. If there is a newly arriving element of the sequence at the current time, for all combinations of candidate elements in the arriving data ,examine if the element is on the path from root to the current node. If the element is not on the path, it means there is a new candidate sequential pattern.If the sequence has already been in the sequence list, the timestamp of

the sequence in the child node should be updated as the new one. Thus, the sequential patterns in the POIs between the old timestamp and the new one can be found. Additionally, for the elements after the new timestamp, appending them to the node having the sequence with the new timestamp is the only way to find up-to-date sequential patterns in the POI beginning at the new timestamp.

After processing common nodes, we have already generated all candidate sequential patterns of all sequences by appending elements to the nodes appearing at the previous timestamps. Then, if the number of sequence IDs in a sequence list is larger

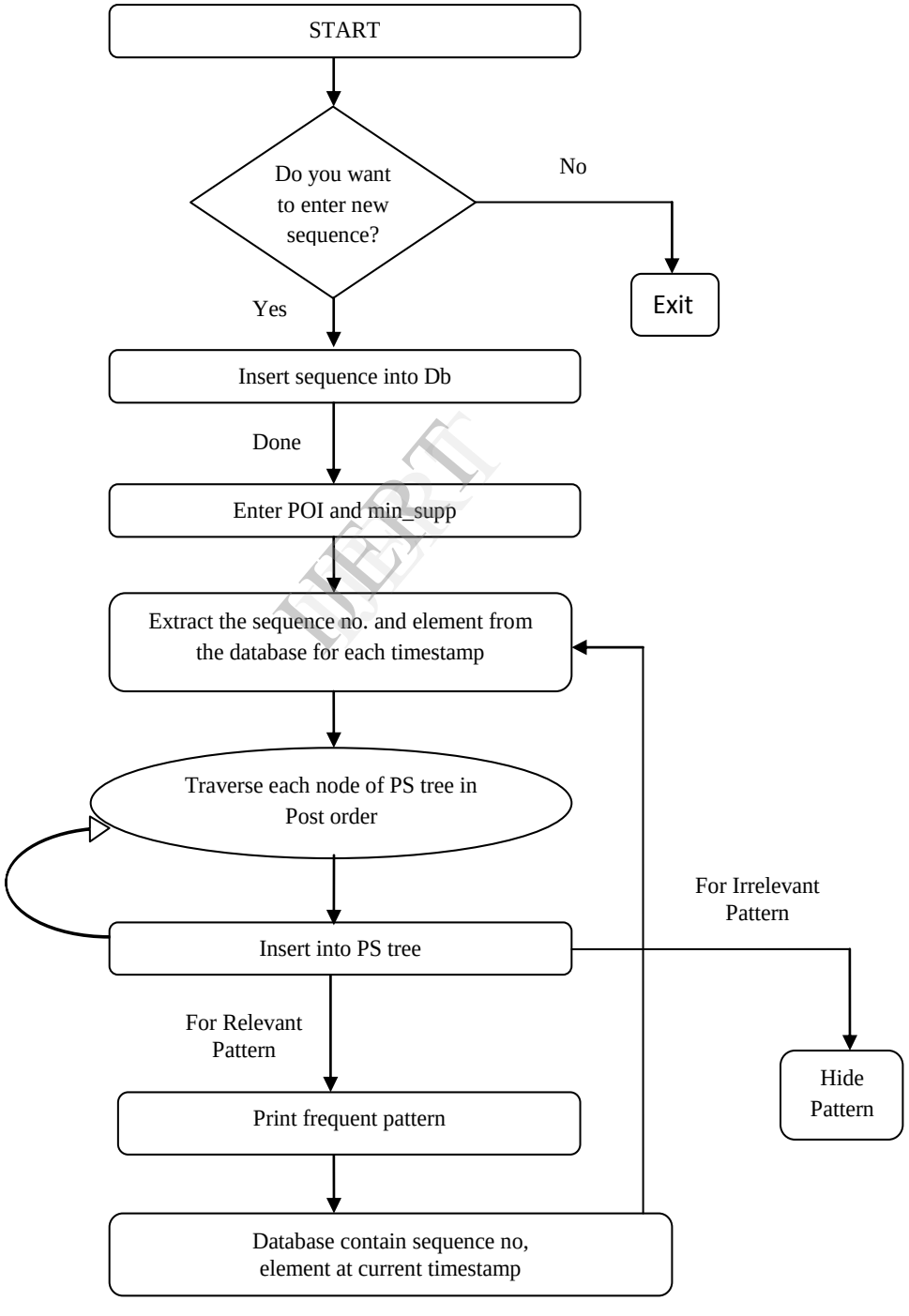

Figure 1: Control Flow Diagram

than the minimum support time's number of sequences in the current POI, the path from root to this node will be printed as a frequent sequential pattern. If the sequence is not relevant one then it hides the entered sequence and do not show any frequent pattern. Then, all the candidate patterns of new sequence are added to database and create new node for such sequence.

10. Advantage of Progressive Mining.

Fast Performance.

Complexity reduced.

Integrity is achieved.

By using PISA, we can progressively discover sequential patterns in defined time period of interest (POI).[13]

Pisa utilizes a progressive sequential tree to efficiently maintain the latest data sequences.It can discover the complete set of up-to-date sequential patterns. [13]

It can also delete obsolete data and patterns accordingly.

No need to worry about identification of newly arriving patterns due to the existence of old data and sequences when sequential patterns will get generate.

11. Application of sequential data mining

1) Mining transactional data

It is possible to mine sequential patterns in sequences of transactions from a store [17].In this case, each sequence represents the transactions from a customer at the store. From this, a sequential pattern mining algorithm could find patterns common to several customers. For example, For example, customers who buy bread and biscuit buy.

2) Mining stock market data

Various stock market analysis on stock data is possible with sequential pattern mining.

2) Mining web logs

Sequential pattern mining can be done on web logs. In this various sequences of weblogs visited by users on website are mined to find sequences of pages that are frequently visited by users. The website could then use these patterns to generate suggestions to the user such as recommended links [16] [17].

4) Software engineering

To find the pattern in source code sequential pattern mining is applied as one of application in software engineering.

5) Mining medical records.

Data about the various medicines, patient record, and disease diagnosis test result are included in medical records. Sequential pattern mining algorithm could be used to find patterns in medical records. Example: Medicine is provided to patient according to their physical strength. To find pattern which medicine is more suitable for which patients [16].

6) Mining educational data

Sequential pattern mining can be used to find patterns in educational data [17]. Mining on result record, Course record for which students are frequently appeared are some example of mining educational data.

12. Conclusion:

Sequential pattern mining with progressive database provide faster performance, reduce complexity and maintain integrity of sequences as compare to static and incremental database. It should consider the most recent items and they are scanned only once. Sequential pattern mining is widely used technique that is applied on various types of database records. Progressive mining algorithm efficiently handles the problem of sequential pattern mining over progressive data. Progressive mining of data can be extended to provide a certain amount of data hiding in this, the actual data is hidden from the user by adding certain fake values, such that they do not affect the original outcome of the algorithm.

References:

[1] B.N. Lakshmi , G.H. Raghunandhan," A Conceptual Overview of Data Mining", Proceedings of the National Conference on Innovations in Emerging Technology, pp.27-32, February 2011

[2] Qi Luo, "Knowledge Discovery and Data Mining," in Proc. Workshop on Knowledge Discovery and Data Mining, Adelaide, SA, 2008, pp 3-5, IEEE.

[3] M.Dunham. "Data Mining – Introductory and Advanced Topics". Pg 185-186. Section 6.7.2. Pearson Education. 2003

[4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence Magazine, pp. 37-54, 1996

[5] Sizu Hou, Xianfei Zhang, "Alarms Association Rules Based on Sequential Pattern Mining Algorithm," In proceedings of the Fifth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 2, pp.556-560, Shandong, 2008.

[6] Jatin D Parmar and Sanjay Garg, "Modified Web Access Pattern (mWAP) Approach for Sequential Pattern Mining", Journal of computer Science, Vol. 6, No.2, pp.46-54, June 2007.

[7] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufman publishers, 2001, ISBN: 1-55860489-8.

[8] Agrawal, R. and Srikant, R. Mining sequential patterns. In Eleventh International Conference on Data Engineering, P. S. Yu and A. S. P. Chen, Eds. IEEE Computer Society Press, Taipei, Taiwan, pp. 3-14, 1995.

[9] Myra, "Web usage mining for Web site evaluation", Communications of the ACM, vol. 43, No. 8, pp. 127–134, 2000

[10] S. Parthasarathy, M.J. Zaki, M. Ogihara, and S. Dwarkadas, "Incremental and Interactive Sequence Mining," Proc. 8th ACM Int'l Conf. Information and Knowledge Management (CIKM '99), pp. 251-258, 1999.

[11] F.Masseglia, P.Poncelet, and M.Teisseire, "Incremental Mining of Sequential Patterns in Large Databases,"Data and Knowledge Eng., vol. 46, pp. 97-121, 2003.

[12] H. Cheng, X. Yan, and J. Han, "INCSPAN: Incremental Mining of Sequential Patterns in Large Database," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '04), pp. 527- 532, 2004

[13] Jen-Wei Huang, Chi-Yao Tseng, Jian-Chih Ou, and Ming-Syan Chen. "A General Model for Sequential Pattern Mining with a Progressive Database", Knowledge And Data Engineering, vol. 20, no. 9, pp. 1153-1167, September 2008.

[14] Shilpa and Sunita Parashar"Static Data Mining Algorithm with Progressive Approach for Mining Knowledge" Global Journal of Business Management and Information Technology. Volume 1, Number 2 (2011), pp. 85-93

[15] A. Balachandran, G.M. Voelker, P. Bahl, and P.V. Rangan, "Characterizing User Behavior and Network Performance in a Public Wireless LAN," Proc. ACM SIGMETRICS Int'l Conf. Measurement and Modeling of Computer Systems(SIGMETRICS '02), pp.195-205, June 2002

[16] Manish Gupta, JiaWei Han "Application of pattern Discovery using Sequential Data Mining" dais. cs. uiuc. edu/manish/pub/guptal11b_apdsdm.pdf

[17] "Applications of sequential pattern mining" Posted by: webmasterphilfv Date: April 02, 2012http://forum.ai-directory.com/read.php?5,501,1052.