Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 1, January -2015

A Survey: On Various Incremental Association Rule Mining Approaches

Darshak I. Sojitra[1], Dr. K. M. Patel[2]

[1]Computer Engineering, RK. University, darshak.patel8999@gmail.com
[2]Computer Engineering, RK. University, kamlesh.patel@rku.ac.in

Abstract — *Data mining has been pervasively used for extracting business intelligence to support business decision making processes. One of the most fundamental and important tasks of data mining is the mining of frequent patterns. In real life, new transactions are continuously added to the database as time advances. This result in; periodic change in correlations and frequent patterns present in database. Incremental Association Rule mining is used to handle this situation. We could use the previous analysis to incrementally mine the frequent itemset from the updated database, the mining process would become more efficient and cost of mining process would be minimized. A survey is done on the different methods of IARM. Different approaches were defined and advantages and disadvantages of them are discussed. Also find out various issues on incremental association rule mining.*

Keywords- *Data mining, Frequent Itemset, Incremental Association Rule Mining, and Promising frequent itemset.*

I. INTRODUCTION

Recent works in the field of databases have been used in business management, government administration, scientific, engineering data management, and many other applications. This explosive growth in data and databases has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge.

During the past few years, data mining has been considered as new techniques and tools for intelligently transforming the processed data into useful information and knowledge. Data mining provides the technique to analyze and convert mass volume of data and/or detect hidden patterns in data into valuable information. Data mining may be applied in an intelligence environment in a number of domains.

Association rule mining, one of the most important and well researched techniques of data mining. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Various association mining techniques and algorithms will be briefly introduced and compared later.

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two subproblems.

- Find those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent or large itemsets.
- Generate association rules from those large itemsets with the constraints of minimal confidence.

Suppose one of the large itemsets is L_k, $L_k = \{I_1, I_2 \dots I_k\}$, association rules with this itemsets are generated in the following way: the first rule is $\{I_1, I_2 \dots I_k\text{-}1\} \Rightarrow \{I_k\}$, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty.

The remaining of the paper is organized as follows. Section 2 reviews related work in incremental association rule mining. Section 3 discusses incremental association rule mining concept. Comparison of incremental association rule mining approaches are provided in Section 4. Finally, Section 5 concludes with a summary of those incremental association rule mining algorithms.

II. PREVIOUS WORK

In 1993, association rule discovery was first proposed by Agrawal et al. in a pilot study in market basket analysis which found the relationship between the buying items in a retail transaction database. Next year, Apriori, the most popular algorithm of association rule mining, was issued. Apriori is normally divided into 2 major steps: finding frequent itemsets (sometimes called large itemsets) and generating rules. After Apriori was revealed, there are many researchers propose algorithms in this field.

Kuldeep and Neeraj introduce enhancedFP which does its work without prefix tree or any other complicated data structure and there is no re-representation of transaction is necessary [2].

Sanjay and Ketan Proposed a graph based approach, this method can be divided in three parts: To make a graph for the given database. In the second part, remove all the non-frequent nodes and readjust all the links, and in the third part frequent itemsets are mined from the pruned graph [3].

Jyoti, Lata and Vijay introduced novel method for incremental discovery of frequent patterns using Main Memory database Management System, is divided into two types of modules; Central coordinator module and Parallel processor module. Central coordinator module divides the available Dataset or increment to Dataset into N parts using horizontal partitions; here N is number of available processors. Parallel processors perform their task to generate frequent patterns and return the result to central coordinator. Finally Central coordinator combines all the results [7].

Nancy P. Lin propose a novel algorithm, called DSPID, which takes full advantage of the information obtained from previous mining results to cut down the cost of finding new sequential patterns in an incremental database [11].

Palak Patel introduce Improved FP-growth algorithm which changed the data structure of FP-Growth algorithm and it is applied on incremental database. So, this algorithm can work better for incremental database and produces frequent items and reduce time rather than other incremental mining algorithm [1].

Anju and Anita introduce new approach of incremental association rule mining for finding frequent itemset and promising frequent itemset based on bucket sort algorithm without scanning old database. This does not scan original database. I.e. without scanning original database it will scan only incremented database. The itemset which are not frequent in original database but it could be frequent when incremented transaction are added to database is called promising frequent itemset [4].

Chetashri, Ketan and Prajakta propose a novel incremental mining scheme with a parallel approach for discovering frequent itemset. This uses parallel approach for counting frequent itemsets using IMBT data structure. The main objective is to parallelize IMBT creation and frequent itemsets counting, to be run on multiple machines at high speed when memory and processor limitations would make it impractical to run the algorithm on a single machine [7].

Prajakta proposes a novel incremental learning algorithm that makes use of a data structure called Item-Itemset (I-Is) tree that is a variation of B+ tree. The created I-Is tree is updated incrementally [8].

Shilpa propose a new algorithm named progressive APRIORI (PAPRIORI) that will work rapidly. This algorithm generates frequent itemsets by means of reading a particular set of transactions at a time while the size of original database is known [10].

Araya and Worapoj introduce improved probability-based incremental association rule discovery using normal approximation to estimate the probability of occurrence of expected frequent itemset. There are 2 main phases: original mining and incremental mining phase [6].

Ratchadaporn and Worapoj propose a new algorithm named promising frequent itemset algorithm. Algorithm uses maximum support count of 1-itemsets obtained from previous mining to estimate infrequent itemsets, called promising itemsets, of an original database that will capable of being frequent itemsets when new transactions are inserted into the original database. Thus, the algorithm can reduce a number of times to scan the original database. As a result, the algorithm has execution time faster than that of previous methods [12].

Kavitha, manjula and kasthuri introduce Efficient Incremental Rule Mining (EIRM) Algorithm. In the only a single scan to the database is needed. Instead of itemsets, the Transaction Identifiers (TIDs) are stored for discovering promising and unpromising item sets and relevant support count are also maintained. This helps to find all frequent patterns of an updated dataset efficiently and reduces execution time [5].

III. INCREMENTAL ASSOCIATION RULE MINING

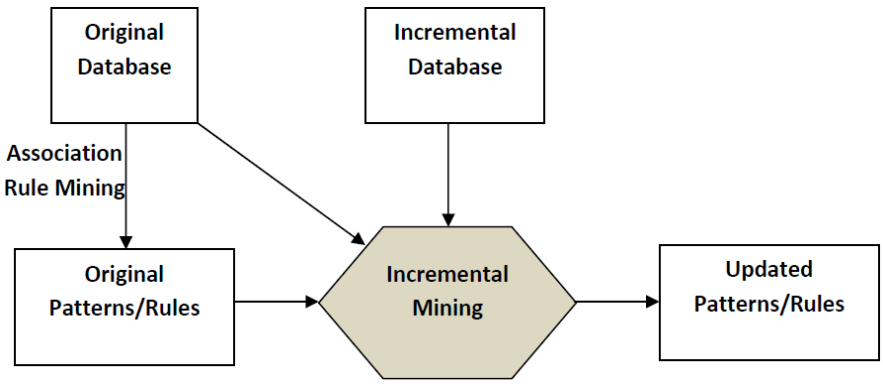

Figure 1. Process of incremental association rule mining

Generally, dynamic database is categorized databases into 2 types: an original database and an increment database. The original database is the old database which old transactions are collected. The increment database is the new database which a new group of transactions are inserted into the original database. When a new increment database

is merged to the original database, the association rule from the previous mining may have been changed. The problem statement for an incremental association rule discovery is normally defined as follows.

Let DB is an original database which is a collection of old transactions. db is an increment database which is a collection of new transactions. Then, UD is an updated database which is the database after merging DB and db together, i.e., UD = DB U db. The number of transactions in a database is called the size of the database. Thus, the size of DB, db and UD are |DB|, |db| and |UD| = |DB|+|db| respectively.

As mentioned to Tsai et al. [13], when an original database and increment database are merged, an itemset, i.e., X, can possibly become to 4 cases:

- Case 1: X is a frequent itemset in both DB and UD
- Case 2: X is a frequent itemset in DB and an infrequent itemset in UD
- Case 3: X is an infrequent itemset in DB and a frequent itemset in UD
- Case 4: X is an infrequent itemset in both DB and UD

From all cases mentioned above, the first two cases are easily discovered frequent itemsets in an updated database since their support count are exactly known. Accordingly, the updating tasks in these 2 cases are negligible tasks. In the fourth case, it does not necessarily keep attention because it does not need to rescan an original database. The most difficult task of these 4 cases is the third case because it needs to rescan an original database to obtain the support count of itemsets in the updating tasks. The rescanning an original database is the really big problem because a lot of I/O operations are required.

IV. PRONES AND CONES OF DIFFERENT INCREMENTAL ASSOCIATION RULE MINING APPROACHES

Table 1. Comparison of various incremental association rule mining approaches

Techniques	Short Review	Prone	Cones
Enhanced FP	– Work without complex data structure. – Each transaction is represented as a simple array of item identifier	– Simple data structure and processing scheme. – Fastest than apriori and FP-growth.	– For large dataset require more memory space
FP Graph based approach	– It seems to be easier to apply incremental concepts. – It gives good results in a reasonable time.	– Remove the problem of memory. – Duplication of the node is avoided. Incremental mining is also possible	– Time taken by the graph based approach is more than the tree based approach
Multi processor approach	– It uses two types of modules: Central Coordinator module and Parallel processor module.	– It also works efficiently in single as well as multi processing environment – Gives better and faster performance	– Data partition depend on available processor
Improved FP-Growth based on incremental database	– It uses the linear list table. First they scan all transactions and count item's support put it into ascending order according to its support. Then check minimum support.	– It generates better output than Simple FP-Growth Algorithm in terms of time. – Reduces the runtime and the main memory consumption.	– Previous linear list table is not updated when old transaction was obsolete.
Discover sequential patterns in incremental database (DSPID)	– Use the information obtained from previous mining results to reduce the cost of finding new sequential patterns in an incremental database. No candidates were generated	– Reduce down the cost of finding new sequential patterns in an incremental database. – Saves a lot of memory unit both in hard disk and RAM	– Array implementation is difficult and multiple times scan is required
Promising frequent itemset	– This method will estimate infrequent itemset of original database which is going to be frequent when new transaction are added.	– Reduced the execution time and memory is also reduced. – No need to add the old database with new coming data	– Consider only incremental data

Parallel approach for counting frequent itemsets using IMBT data structure	– It uses a tree structure called IMBT to enumerate the support count of each itemset in an efficient way after the transactions are added or deleted.	– More efficient than existing non-parallel incremental methods such as Apriori and FP-tree methods.	– Faces the problem of limited memory space.
Item-Itemset (I-Is) Trees	– Initially I-Is tree is created from the original data to allow searching of frequent items based on the threshold values. The created I-Is tree is updated incrementally.	– Searching time can be reduced and memory requirement can be handling.	– Child node are depend on the frequency of the root Node
Apriori algorithm with progressive approach (PAPRIORI)	– Initially all 1-itemsets are set as estimated itemsets. Now read database with K transactions at a time. Calculate large itemsets until number of transactions read is less than total number of transactions in database with increase in transactions read with a parameter K.	– Works effectively and efficiently as compared to APRIORI algorithm	– Execution time depends on datasets, minimum support value and value of K. – Difficult to used for incremental dataset

V. CONCLUSION

In this paper we briefly reviewed some algorithms for generating association rules from a dataset. Number of dataset scans and the number of candidate itemsets generated are the two major challenges of the rule mining problem. Datasets used in most of the rule mining algorithms are assumed to be static. But in reality, they may be updated time to time. This incremental behavior of the dataset becomes another challenge to the rule mining problem. Several algorithms attending this issue of the rule mining problem are presented in this paper. Most of the algorithms try to reduce the number of scans of database and maintain the association rules efficiently.

REFERENCES

[1] Palak Patel, "ASSOCIATION RULE MINING USING IMPROVED FP-GROWTH ALGORITHM", International Journal For Technological Research In Engineering Volume 1, Issue 10, June-2014
[2] Kuldeep Singh Malik and Neeraj Raheja, "IMPROVING PERFORMANCE OF FREQUENT ITEMSET ALGORITHM", IJREAS Volume 3, Issue 3 (March 2013)
[3] Sanjay Patel and Dr. Ketan Kotecha, "Incremental Frequent Pattern Mining using Graph based approach", International Journal of Computers Technology, Volume 4 No. 2, March-April, 2013
[4] Ms. Anju k.kakkad and Ms. Anita Zala, "Incremental Association Rule Mining by Modified Approach of Promising Frequent Itemset Algorithm Based on Bucket Sort Approach", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 11, November 2013
[5] Kavitha j.k, manjula d and kasthuri bha j.k, "Effective and Efficient rule mining technique for incremental dataset", Journal of Theoretical and Applied Information Technology, Vol. 57 No.3, 30th November 2013.
[6] Araya Ariya and Worapoj Kreesuradej, "Probability-Based Incremental Association Rule Discovery Using the Normal Approximation" , IEEE IRI 2013, August 14-16, 2013
[7] Jyoti Jadhav, Lata Ragha and Vijay Katkar, "Incremental Frequent Pattern Mining", International Journal of Engineering and Advanced Technology (IJEAT),Volume-1, Issue-6, August 2012
[8] Mrs. Chetashri Bhadane, Dr. Ketan Shah and Mrs. Prajakta Vispute,"An Efficient Parallel Approach for Frequent Itemset Mining of Incremental Data", International Journal of Scientific Engineering Research, Volume 3, Issue 2, February -2012
[9] Mrs. Prajakta Vispute and Prof. Dr. S. S. Sane ,"Incremental Learning Algorithm for association Incremental Learning Algorithm for association rule mining", International Journal of Scientific Engineering Research, Volume 3, Issue 11, November-2012
[10] Shilpa and Sunita Parashar, "Performance Analysis of Apriori Algorithm with Progressive Approach for Mining Data", International Journal of Computer Applications, Volume 31 No.1, October 2011
[11] Nancy P. Lin, Wei-Hua Hao, Hung-Jen Chen, Hao-En and Chueh, Chung-I Chang ,"Discover Sequential Patterns in Incremental Database", INTERNATIONAL JOURNAL OF COMPUTERS Issue 4, Volume 1, 2007
[12] Ratchadaporn Amornchewin and Worapoj Kreesuradej, "Incremental Association Rule Mining Using Promising Frequent Itemset Algorithm" , IEEE 2007
[13] T P Hong, C Y Wang, Y H Tao, "A new incremental data minting algorithm using pre-large itemsets," Journal of Intelligent Data Analysis, Vol.5, No.2, 2001, pp.111-129.

