Design and Development of an Efficient Algorithm of Sequential Pattern Mining based on Apriori Algorithm using Association Rules

[1]Lutfi Habiba and [2]Shahida Rafique
[1]*Department of computer Science and Engineering*
[2]*Institute of Science and Technology, National University, Bangladesh*
Corresponding author email- lutfihabiba@gmail.com

Abstract

An efficient algorithm of sequential pattern mining has been designed and developed. The sequential Pattern Mining is based on Apriori Algorithm. Association Rules has been used to generate the strong rules for extracting valuable information from large database. Many approaches are proposed in past to improve Apriori but the core concept of the algorithm is same i.e. support and confidence of itemsets and previous studies finds that Apriori is insufficient due to many scans on database. It improves Apriori algorithm efficiency by reducing the database size as well as reducing the time wasted on scanning the transaction. It has been also found that useful sequences occur frequently in database. These sequences are used in finding users' purchasing behavior in retail Industries, User's access sequences to access web pages, to identify the sequences repeatedly occur and responsible for particular disease etc. The current state-of-the-art methods have not succeeded to produce sequences for large database with minimum database scanning. The algorithm can be used to produce the sequences in large database by reducing database size and time of the items of transactions using association rules.

Keywords: Data Mining, Sequential Pattern Mining, Association Rules, Apriori, Frequent pattern, support, confidence.

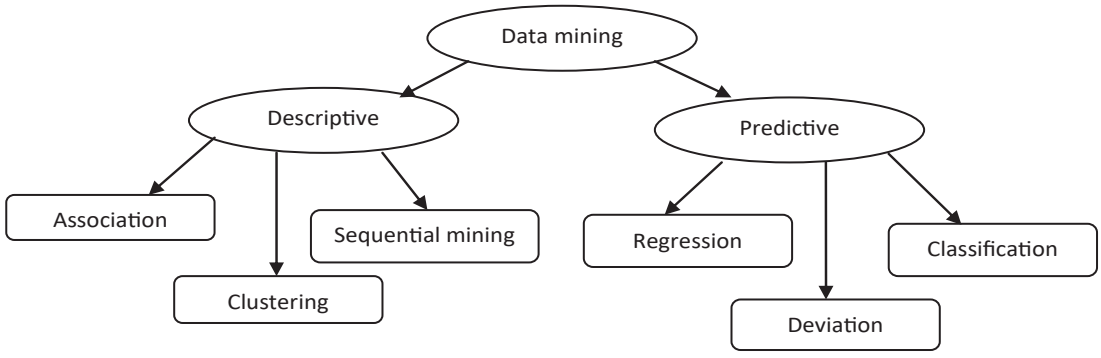
1. INTRODUCTION

Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. Sequential pattern mining is a special case of data mining. There has been a lot of work in the field of data mining about sequential pattern mining. The goal of sequential pattern mining is to discover useful, novel and unexpected patterns in databases [1].

In general, data mining tasks can be classified into two categories: Figure 1 presents the task categories of data mining

Descriptive mining: It is the process of drawing the essential characteristics or general properties of the data in the database. Clustering, Association and Sequential mining are one of the descriptive mining techniques.

Predictive mining: This is the process of inferring sequences form data to make predictions. Classification, Regression and Deviation detection are predictive mining techniques.



2. DATA MINING AND KNOWLEDGE DISCOVERY IN DATABASES

The last decade has experienced a revolution in information availability and exchange via the internet. In the same spirit, more and more businesses and organizations began to collect data related to their own operations. While the database technologists have been seeking efficient means of storing, retrieving and manipulating data, the machine learning community has focused on developing techniques for learning and acquiring knowledge from the data. At times the data can be considered to be a gold mine for strategic planning for research and development in this area which is often referred to as Data Mining (DM) and Knowledge Discovery in Databases (KDD).The formal and complete analysis process is called knowledge discovery from databases (KDD). KDD establishes the main procedures for transforming data into knowledge. The KDD process follows the steps indicated in figure 2 [2]: collection of a target dataset, data warehousing, transformation of the data into adequate forms for the DM process, selection of a DM tool, relationship identification of DM (classes, clusters, associations), interpretation of results, and consolidation of discovered knowledge.

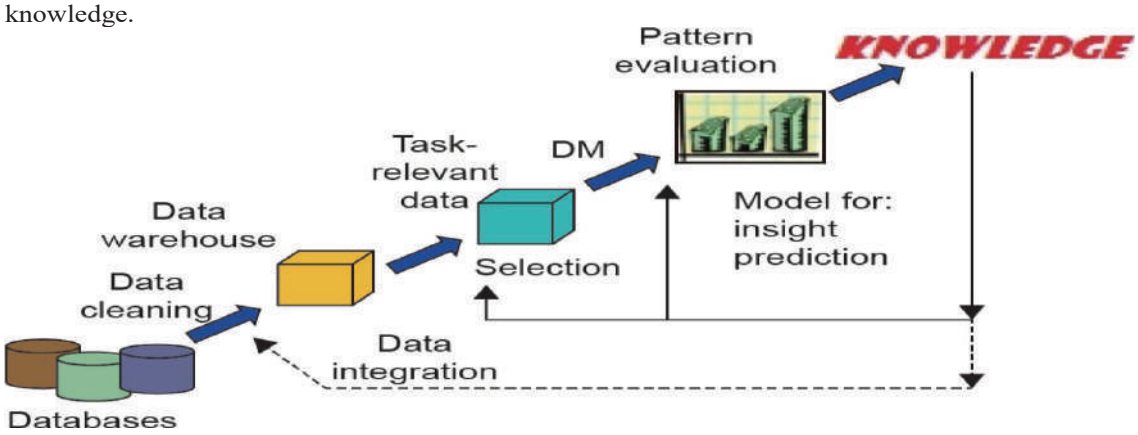


Figure 2: Knowledge Discovery in Databases [2]

3. LITERATURE SURVEY

Several techniques are used to extract sequential pattern mining from web logs. The main techniques can be categorized into Apriori-based, pattern-growth and early-pruning techniques. Apriori-based algorithms are deemed slow and have large search space, while pattern-growth algorithms have been tested extensively on mining the web log and found to be fast Early-pruning techniques have success stories with web access sequences stored in dense databases. An existing sequential pattern mining algorithms are provided based on the above three techniques described below:

3.1Apriori-based Techniques

The first and simplest family of sequential sequence mining algorithms is Apriori based algorithms and their main characteristic is that they use Apriori principle [3].

3.2 Tree-based Techniques

A faster and more efficient candidate production can be attained by using a treelike structure.

3.3 Lattice-based Techniques

Lattice structure was another class of sequential sequence mining algorithms was proposed a lattice based method to enumerate the candidate sequences efficiently

4. ASSOCIATION RULES MINING

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. Generally association rule is applied on the large amount of data. For example, the association technique is used

Table 1:-Market Basket Data Table

TID	Items
10	Butter,Milk,Eggs
20	Tea,Milk,Sugar
30	Butter, Tea,Milk,Sugar
40	Tea, Sugar

Association Rule
An implication expression of the form
Tea → Milk, where Tea and Milk are disjoint itemsets

Support (Tea →Milk) =

$$\frac{\text{No. of transactions containing Tea \& Milk}}{\text{No. of total transactions}}$$

Confidence (Tea→ Milk) =

$$\frac{\text{No. of transactions containing Tea \& Milk}}{\text{No. of transactions containing Tea}}$$

5. EXISTING SYSTEM

An important data mining problem is to design algorithm for discovering hidden patterns in sequences. There have been a lot of research on this topic in the field of data mining and various algorithms have been proposed.

5.1 Apriori Algorithm

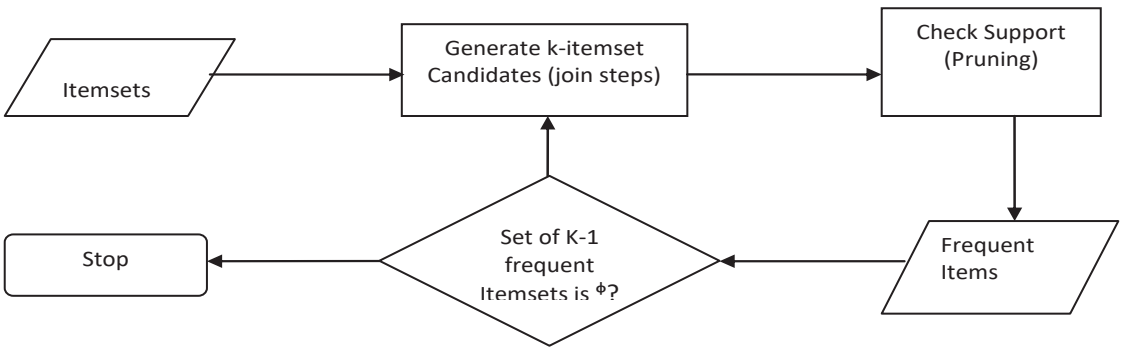


Figure 3. Flow Chart of Apriori Algorithm

The algorithm [4] uses a level-wise search, where k-itemsets are used to explore (K+1)-itemsets. In this algorithm, frequent subsets are extended one item at a time (this step is known as candidate generation process).Then groups of candidates are tested against the data. It identifies the frequent individual items in the database and extends them to large and larger item sets as long as those itemsets appear sufficiently often in the database.

5.2 Experimental Demonstration of Apriori Algorithm

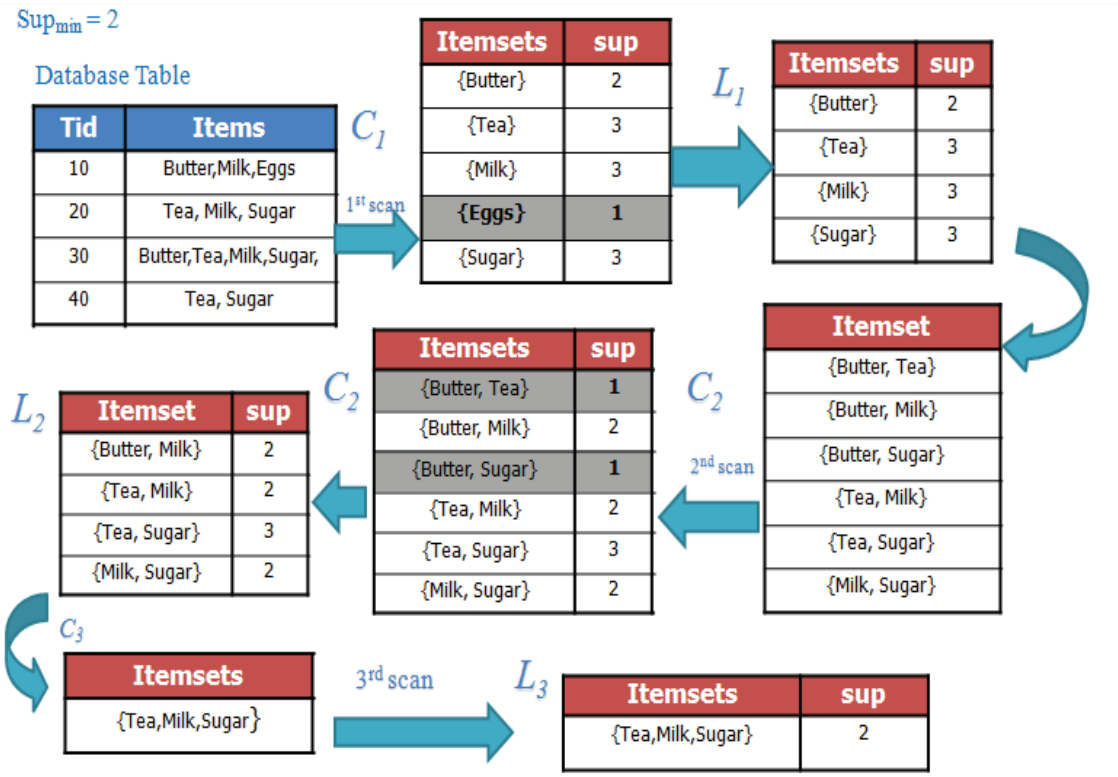


Figure 4.Steps of Apriori Algorithm

From this example several purchase patterns can be observed. For example:

The most popular transaction is of {Tea,Milk,Sugar}
Another popular transaction was of Butter, Milk and other groceries.
If someone buys Tea, he is likely to have bought Milk as well

6. PROPOSED ALGORITHM

In this approach to improve existing algorithm efficiency, it focus on reducing the time consumed for C_k generation. In the process to find frequent item sets, first size of a transaction is found for each transaction in Database and maintained. To find L2 from C2, instead of scanning complete database and all transactions, we remove transaction where Size_of_itemset_Transaction< k (where k is 2, 3…) this helps in reducing the time to scan the infrequent transactions from the database. To generate C3(x, y,z), L3 and so on is generated repeating above steps until no frequent items sets can be discovered.

6.1 Proposed System Flow Chart

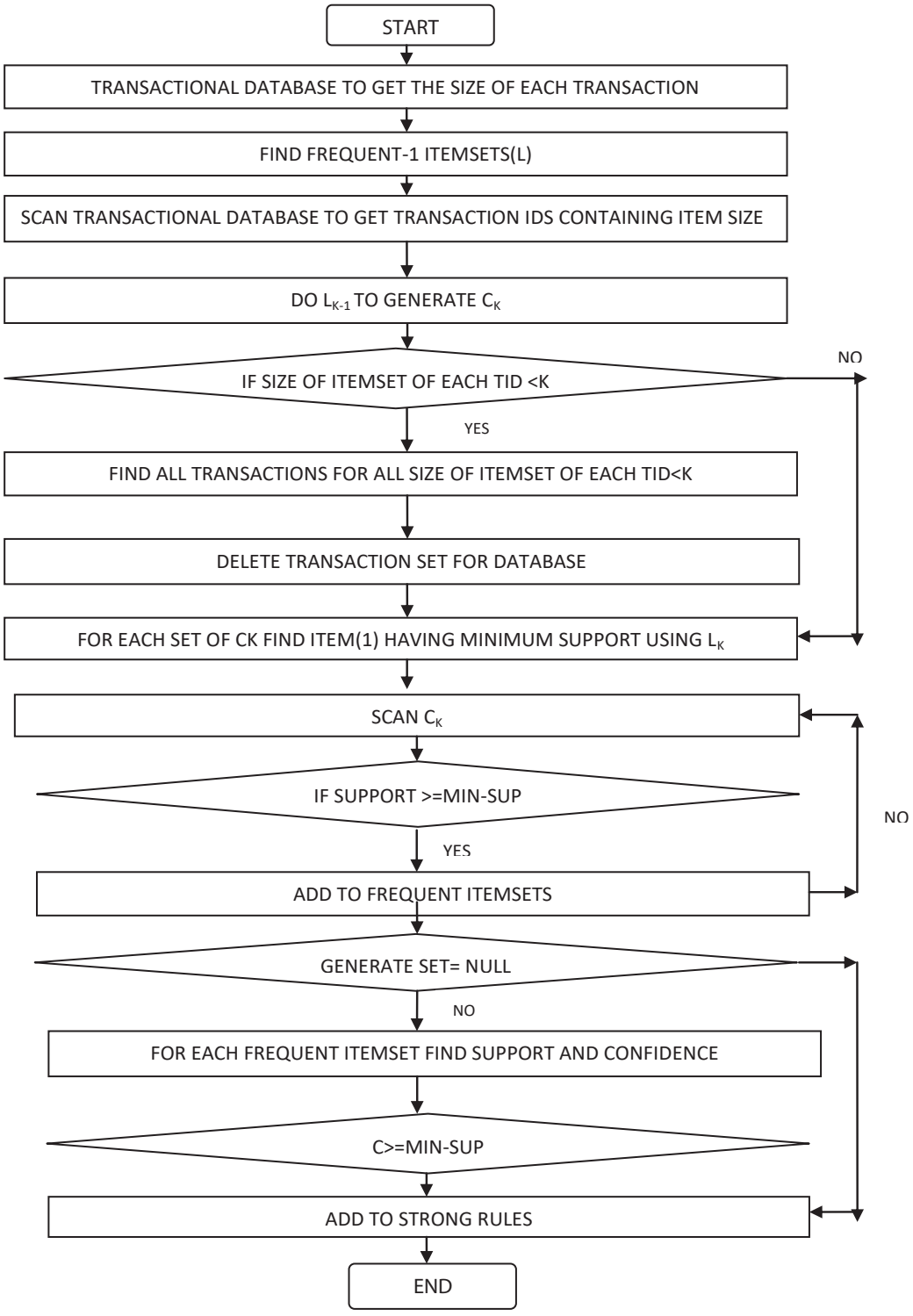


Figure 5: Proposed System Flow Chart

6.2. Experimental Demonstration of Proposed Algorithm

Proposed algorithm demonstration is given below. A database of 10 transactions is considered and size of each transaction is calculated and stored as transaction size. In first iteration, C1 is generated by simple scanning the database to count the number of transactions of each itemset. Min_sup for this example is set as 3. So to generate L1, itemsets having transaction count equal to or greater than 3 are considered as frequent and included in L1. Steps are explained in following figure.

Input: Transactions Database, Minimum support, Min_sup

Output L_k: Frequent itemsets in Database

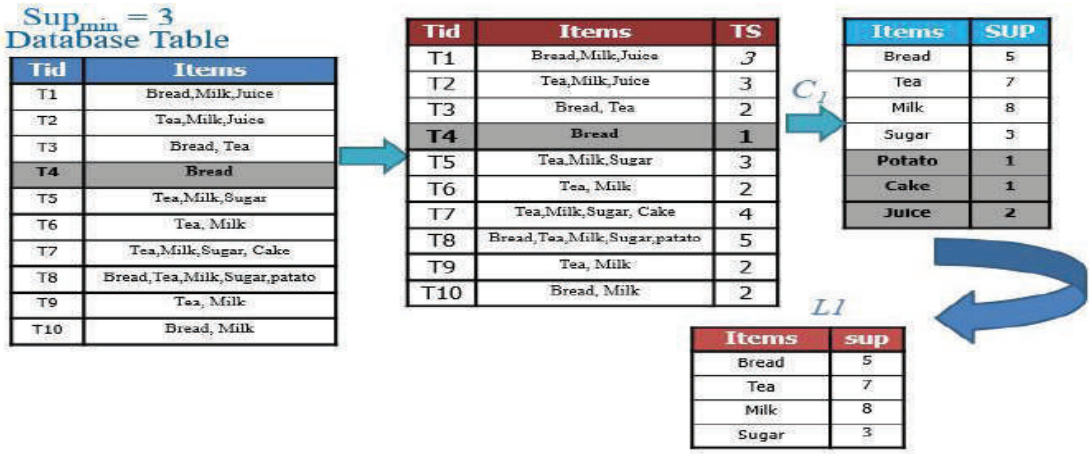


Figure 6: Steps to generate C1 and L1

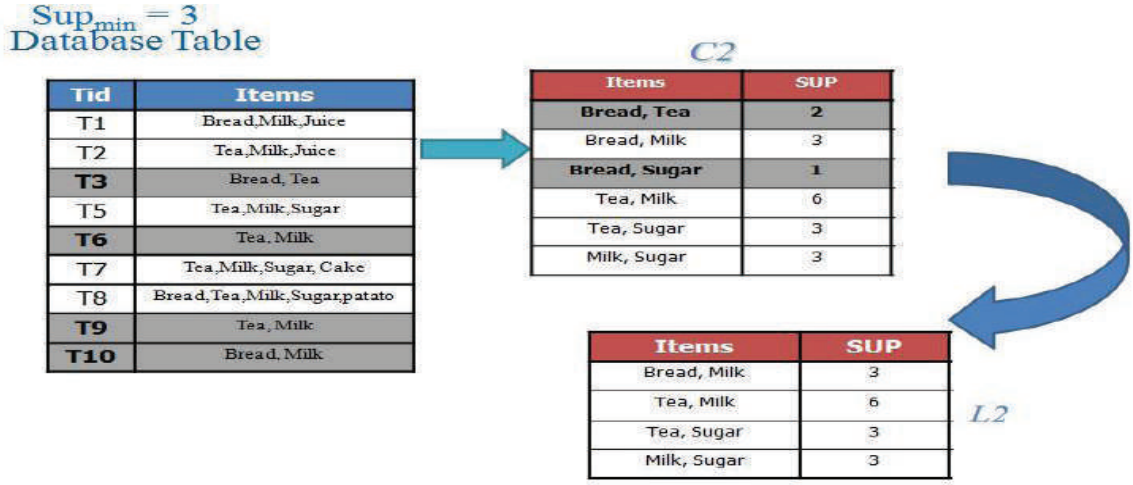


Fig 7: Steps to generate C2 and L2

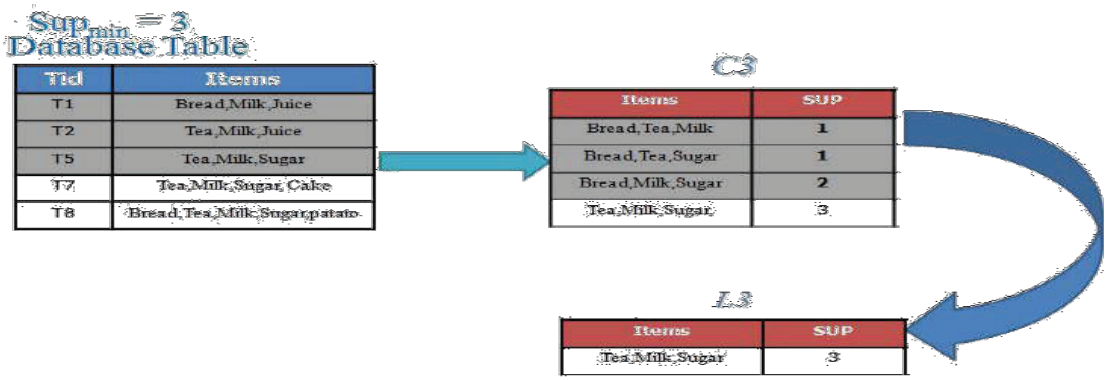


Figure 8: Steps to generate C3and L3

Items	SUP
Tea,Milk,Sugar	3

Association rule	Support	Confidence	Confidence%
Tea^Milk-->Sugar	3	3/5=.60	60%
Tea^Sugar-->Milk	3	3/3=1	100%
Milk^Sugar-->Tea	3	3/3=1	100%

Figure 9: Strong Association Rules

7. SOFTWARE DESIGN CONSIDERATIONS

JavaScript language is used for the system development. JavaScript (often shortened to **JS**) is a lightweight, interpreted, object-oriented language with first, and is best known as the scripting language for Web pages, but it 'used as well. It is prototype, multi-paradigm scripting language that is dynamic, and supports object-oriented, imperative, and functional programming styles. JavaScript is an easy to learn and also powerful scripting language, widely used for controlling web page behavior.

7.1 Proposed System Implementation and Screenshot

The Proposed algorithm for finding large itemsets and generating association rules using large itemsets are illustrated in this Section. Enter a set of items separated by comma and the number of transactions you wish to have in the input database. Then press Generate database button to generate a random database with items that you entered. Then press the Proposed button to see the algorithm in action; a set of large itemsets and association rules will be generated based on the given support threshold and confidence threshold. The source code of the system is implemented using JavaScript language. Here some of my system screenshot is given below.

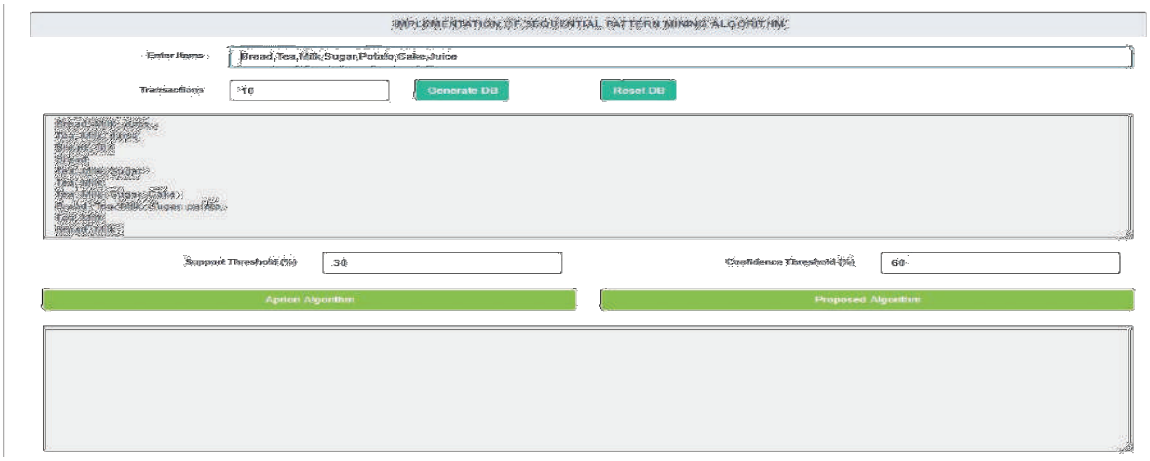


Figure 10: Database Generation

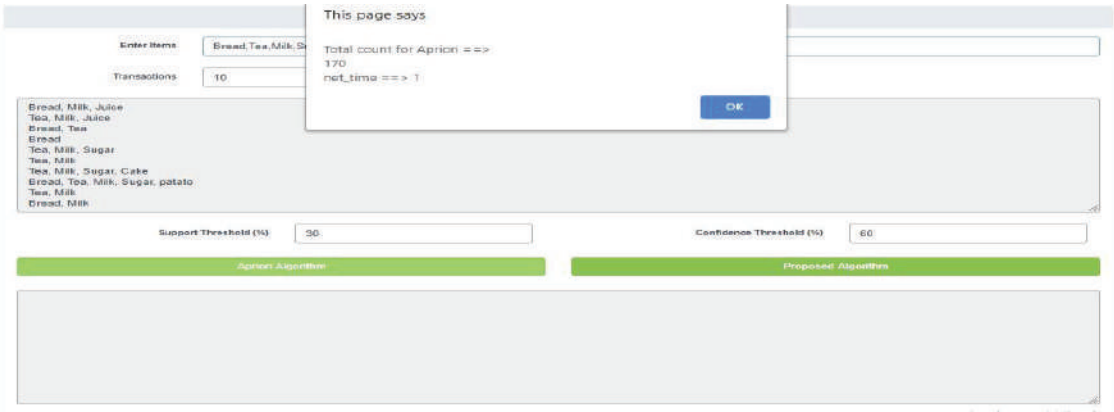


Figure 11: Database Scanning Count (Existing)

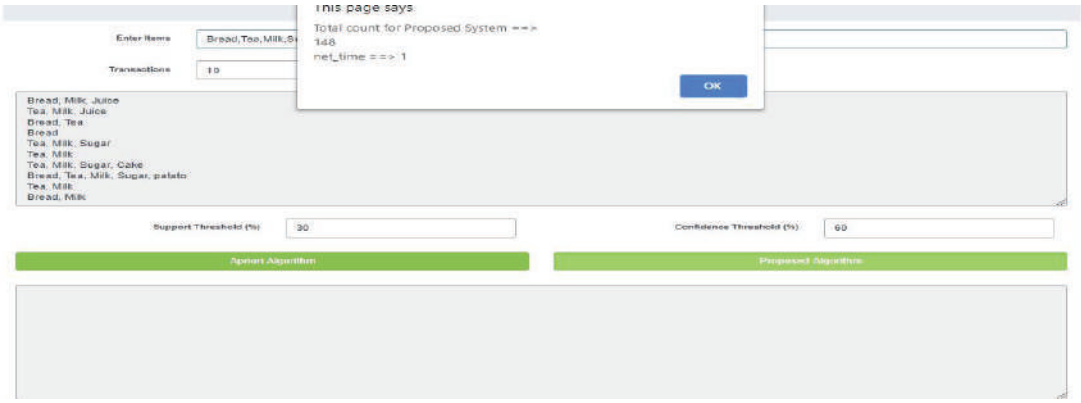


Figure 12: Database Scanning Count (Proposed)

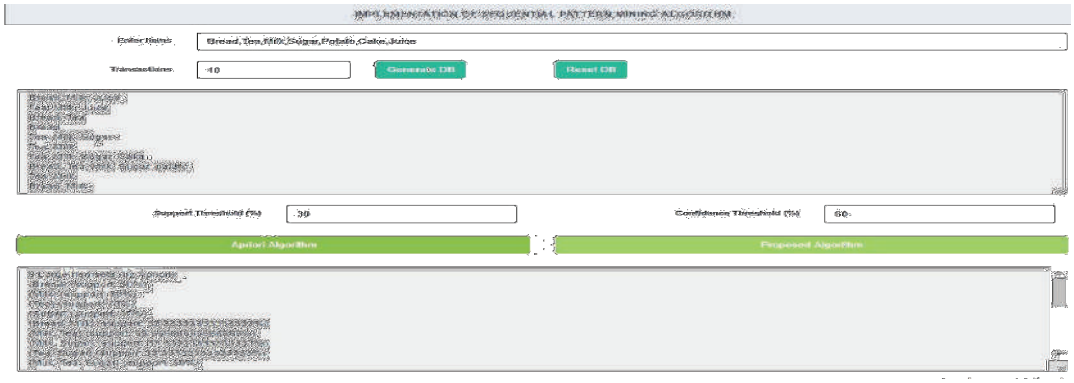


Figure 13: Strong Rules Generation

7.2 Performance Analysis

To evaluate the performance of the algorithms over a large range of data characteristics, we generated data set for transactions. It was tested on data base means the size of the items for 7 items and the number of 10 transactions or we can generate larger database through the system. The proposed system counted the number of transaction that are scanned to find L_1,L_2,L_3 for the given data sets. The following figure shows the differences between the existing system and the proposed system.

Table 2: Performance Evaluation

K	Apriori	Proposed
Category 1	70	70
Category 2	60	54
Category 3	40	24
Total Scanning	**170**	**148**

For K=1 the number of transaction scanned is same for the both system. But the increase in k, count of transactions scanned decrease. The following figure depicted the real scenario.

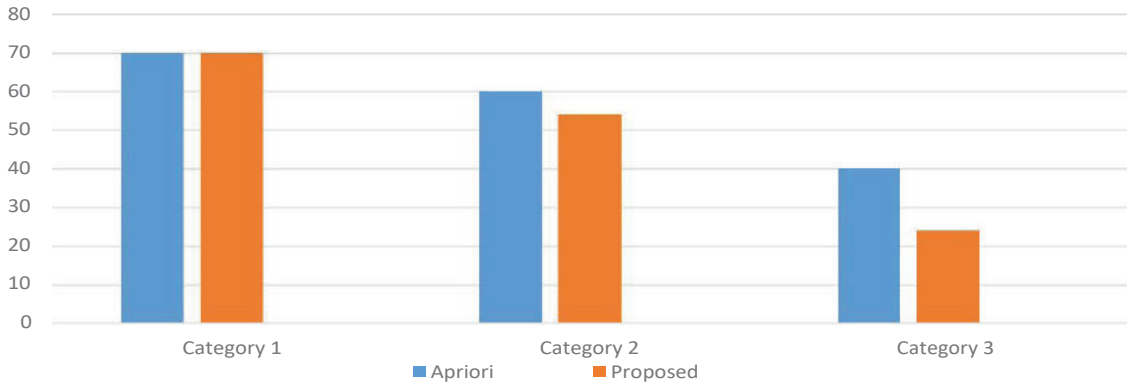


Figure 14: Comparative analysis of algorithm performance based on total scanning

CONCLUSION

The proposed algorithm generate the sequential sequences by proposed algorithm in very efficient way. With the observation and work it could conclude that proposed algorithm provides better performance compared to all earlier algorithms produced for sequential sequences. The empirical analysis and test results state that proposed algorithm outperforms the state-of-the-art methods because of using sequence generator table. The sequence generator table saves the time during execution and decrease the memory usage during execution. Further, it may improve the performance of the algorithm in future as extension the work .It may be more efficient and effective by considering time and memory. Future work focuses on to find a solution to reduce the overhead to maintain new database.

REFERENCES

[1]. R. Agrawal and R. Srikant, "Mining Sequential Patterns", Proc. 1995 Int"l Conf.Data Eng. (ICDE "95), Pages3-14, Mar. 1995.

[2].NIZAR R. MABROUKEH and C. I. EZEIFE, ‖A Taxonomy of Sequential Pattern Mining Algorithms‖, ACM Computing Surveys, Vol. 43, No. 1, Article 3, Publication date: November 2010.

[3]. R Agrawal, R Srikant, "Fast Algorithm for Mining Association Rules", Proc. 20th Int"l Conf. Very Large Data Bases, VLDB, Pages 487-499, 1994.

[4]. S.Rao,R Gupta ,"Implementing Improved Algorithm Over APRIORI Data Mining Association on Rule Algorithm ",International Journal of Computer Science And Technology PP. 489-493,Mar.2012.