The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2995 - 3002

Efficient Pattern Mining Using Integrated Range of Item Sets

Srujan V.
Dept of CSE
Bangalore Institute of Technology
Bangalore, India
srujanv010@gmail.com
Suhaas Rao S.
Dept of CSE
Bangalore Institute of Technology
Bangalore, India
suhaas0000@gmail.com
***Yeshwanth S. P.**
Dept of CSE
Bangalore Institute of Technology
Bangalore, India
*yashwanthsp76@gmail.com
Suneetha K. R.
Dept of CSE
Bangalore Institute of Technology
Bangalore, India
suneethakr@bangalore.edu.in

Abstract—Association rule mining is one of the well-known data mining techniques to identify user interest and their behavioural pattern. It's known as market basket analysis used in all kinds of business-oriented fields to understand customer's interest. The Apriori is used to find frequent by analysing transaction database and to generate association rules. The existing Apriori algorithm suffers with usage of more storage and execution time as it needs to scan the main data base frequently each time while generation of candidate item sets. The proposed algorithm reduces time complexity, input-output load, and memory utilisation by eliminating some of the duplicate combinations which reduces number of scans required to achieve accurate and reliable results with high efficiency.
Keywords—Data mining, Frequent pattern mining, Apriori algorithm

I. Introduction

We live in an era of data deluge, which makes it necessary to classify and filter it through data mining which provides patterns and connections in vast amounts of data from many sources.

Data mining, often referred to as Knowledge Discovery Data (KDD) [13], is the act of extracting knowledge from data to provide recommendations to business problems through analysation. Different types of data mining techniques [14] like, Clustering, Classification, Genetic Algorithms, Regression, Association rule learning, Anomaly detection, Artificial

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2995 - 3002

Neural Network Classification are available to mine the data. This paper focuses on Association rule mining technique and proposes modified methodology to gain performance.

A. Association Rule mining

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently an itemset occurs in a transaction. Association rule mining is a procedure which aims to observe frequently occurring patterns, correlations, or associations from datasets found in various kinds of databases such as relational databases, transactional databases is also known as market basket analysis.

B. Associate Rule mining algorithms

Three different types of algorithms that can be used to generate associate rules in data mining are,

- Apriori Algorithm identifies frequent patterns from the transactional database.
- Eclat Algorithm is an enhanced version and more useful variation compared to Apriori algorithm.
- FP-growth Algorithm *is* very helpful for recognizing regular patterns without the necessity for candidate generation.

This paper works with original concept of Apriori algorithm and proposes modified approach to improve the efficiency in terms of memory usage and time consumption.

Apriori algorithm [11] used for locating frequent itemset in a dataset. The algorithm is known as Apriori since it makes use of frequent item qualities that are known in advance. It employs an iterative technique or level-wise search to identify the k+1 itemset using the k-frequent itemset as a starting point. The Apriori technique has limitations like it is slow for larger data sets and takes a long time to execute. An important attribute known as the Apriori property is used to increase the effectiveness of level-wise generation of frequent item sets by minimising the search space. Towards this direction the paper provides details of performance efficiency by reducing the number of scans to generate association rules which int turn reduces memory usage and execution.

Layout of the paper is: Section II discusses Related Work, III contains the proposed algorithm along with its description, IV provides Pseudo code, V presents Results and Analysis, VI contains Conclusion, followed by references.

II. Related Works

This section gives literature survey on contribution towards association rule mining with different methodologies for extracting frequent patterns from transactional data base.

The proposed algorithm in [1] uses an estimation procedure to decide the number of passes made and it uses a pruning technique. This algorithm also uses buffer management to handle all the item sets.

In [2], authors minimize repeated scans of transactions by splitting the candidate set and utilising bit maps to remove un-necessary computations. The generation of the hash tree may be successfully parallelized by IDD (Intelligent data distribution) technique, making it scalable in terms of growing the candidate list.

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2995 - 3002

For effective large itemset generation, J. Han et.al [3] have designed the DHP (direct hashing and pruning) approach that includes three key characteristics: effective reduction of transaction database size, efficient creation for big item sets, and the choice to reduce the number of database scans necessary. By adopting a hashing methodology, DHP is particularly effective at producing candidate large item sets. The quantity of candidate large item sets produced by DHP is orders of less magnitude, significantly increasing the process's performance bottleneck.

The study in [4] demonstrates that mining partial periodicity only requires two scans of the time series database. The efficiency of the proposed strategy is demonstrated by the performance study. A wide range of applications are covered by partial periodicity since it is less limiting than full periodicity and merely correlates periodic behaviour.

The authors Jiawei Han et.al [5] developed an efficient FP-trees method for frequent pattern identification. The innovative frequent pattern tree structure is an extended pre fix tree structure for storing compressed, important information about frequent patterns.

Authors [6] have introduced the concept of UserIDs to produce the next candidate set. The database is divided into blocks using dynamic itemset technology in this work.

Authors D. Sun et.al [7] incorporated the technologies of numerous enhanced Apriori algorithms. The proposed algorithm does not create candidate item sets. It reduces the amount of time and space needed for frequent item searches.

A new approach based on speeding up candidate set scanning has been proposed in this paper [8]. The hash structures are used to store candidate sets. The Lk-1 itemset is only scanned once in the proposed technique when producing the collection of candidate itemsets Ck. The outcomes show that the modified algorithm is more efficient than the traditional algorithm which reduces the time of scanning candidate sets. Hash structure is used for storage of candidate sets.

In [9] pruned optimization technique is developed and with this approach, the creation of frequently occurring itemsets is decreased, and transaction reduction is employed to condense database transactions. In compared to the Apriori algorithm, the suggested BE-Apriori method is more efficient by reducing the execution time in turn decreases system overhead.

The original Apriori algorithm is designed using a bottom-up approach but the authors in [10] used a top-down approach to minimize unnecessary rules. The benefit of this approach reduces the of number of database scans.

Yubo Jia et.al [11], developed an enhanced technique based on combination of dynamic itemises counting and data division. The primary issues with the traditional Apriori technique have been resolved by proposed algorithm in which the transactional database is separated into separate, non –intersecting sections as part of data division. The main advantages of the proposed work are reduction of I/O load and storage space.

A property called Size of Transaction (SOT), which contains the number of items in each unique transaction in the database is included in [12]. The enhanced Apriori technique is suggested in this research which minimizes size of the candidate set of k itemsets, also lowers Input-Output consumption by reducing the number of transactions in the database.

In [13] the authors have covered alternative methods on Apriori algorithm.

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2995 - 3002

The enhanced method put out in this study [14] operates in two stages. B matrix, a necessary compressed data structure, is built in the first phase before being employed in the second phase to create regular itemsets. As a result, I/O costs are significantly decreased, and irregular itemsets are also avoided. In comparison to traditional Apriori technique, the revised algorithm has superior efficiency and reduces both temporal and spatial complexity.

The authors in [15] implemented Improved version of Apriori algorithm to decrease number of operations to be examine for candidate item sets, to save searching time for potential itemsets. The author claims that the proposed algorithm takes 67.38% less time to create candidate support counts compared to original Apriori algorithm.

The basic Apriori algorithm is modified in [16] to generate frequent patterns and focuses on bringing down the number 2^N to total number of edges formed by the graph. A combined concepts of graphs and logic of Apriori algorithm is used to avoid repeated scanning of data base in which frequent web pages identified in the first pass and relation between the frequent web pages in the second pass and then the graph is mined.

Literature survey shows that the work carried so for will not provide efficient methodology for the issues of Apriori algorithm, hence the objective of this work is to provide better approach in order to reduce number of scans and execution time.

III. Working Methodology of Proposed Apriori Algorithm

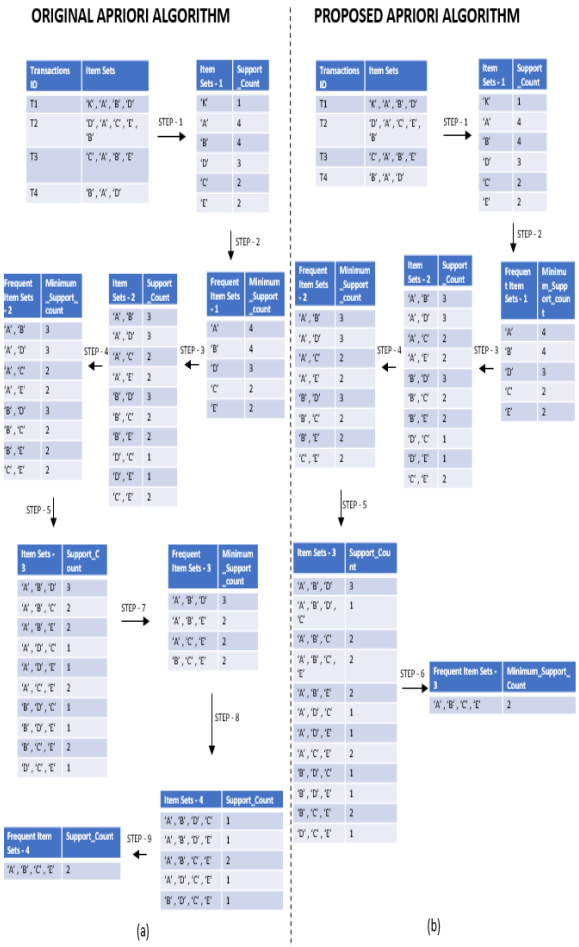

Fig.1. Comparison of the proposed and original Apriori algorithm.

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2995 - 3002

The Proposed Apriori algorithm has the same working procedure till Step 4 (Fig. 1) with support count and confidence metrics. The variation form step - 5 onwards is depicted in Fig 1.Hence, ‘A’, ‘B’, ‘C’, ‘E’ is the most Frequent item set.

In this example, all the item sets generated further (Step 5 – Fig. 1.a) in the original Apriori algorithm (Fig. 1.a) are integrated in the proposed Apriori algorithm (Fig. 1.b) which reduces the traversal of the input database 3 times by avoiding making some unnecessary item sets in Item Sets – 3. (Item Sets – 4 in original Apriori algorithm.)

Consequently, it has been observed that, for this particular example the proposed Apriori algorithm scans the database 28 times totally, whereas the original Apriori algorithm scans the database a total of 31 times. This avoids making a few combinations of item sets.

IV. Pseudo Code of Proposed Methodology

This section provides pseudo code for the proposed methodology.

Input: Database – D, Minimum support count - min_sup.

Output: Frequent item set – T.

```
L1 = [Large 1 itemsets]
    C1[length(L1)] = []
pos = 0
for I in L1
     C1[pos] = i.count
     pos++
L2 = []
for I in L1
    if(i.count<min_sup)
        L1.delete(i)
        L2.append(i)
If(Ti == supersets_of_L2)
    D1.delete(Ti)
T=[]
for I in D1
    T = subsets_transaction_t_of_length_2(Ti)
    for I in T
        for j in D1
            if(I in j )
                i.count++
    If(i.count<min_sup)
        T.delete(i)
maxcount = 2
for I = 0 to length(T)
     Do{
       for j = 0 to length(L1){
            if(L1[j] == itemset[pos])
                Index = j+1
```

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2995 - 3002

```
29.                 T[i].append(L1[index])
30.                 for k = 0 to n
31.                     if (I in D1[k])
32.                         Count++
33.                 if(i.count >= min_sup)
34.                     T.append(i)
35.                     if(length(i)>maxcount)
36.                         maxcount++
37.                 Pos++
38.               }while(index<length(L1))
39.     for I in T
40.         if(i.length != maxcount)
41.             T.delete(i)
```

Fig. 2. Pseudo code for the proposed Apriori algorithm.

V. Performance Analysis

For performance analysis experimental benchmark dataset 6M-0K-99K is taken from Kaggle and data of Global C2C Fashion Store User Behavioral Analysis. The graph shown in Fig 3 validates the proposed work which concludes that time and memory expenses are lesser as compared to the original Apriori algorithm.

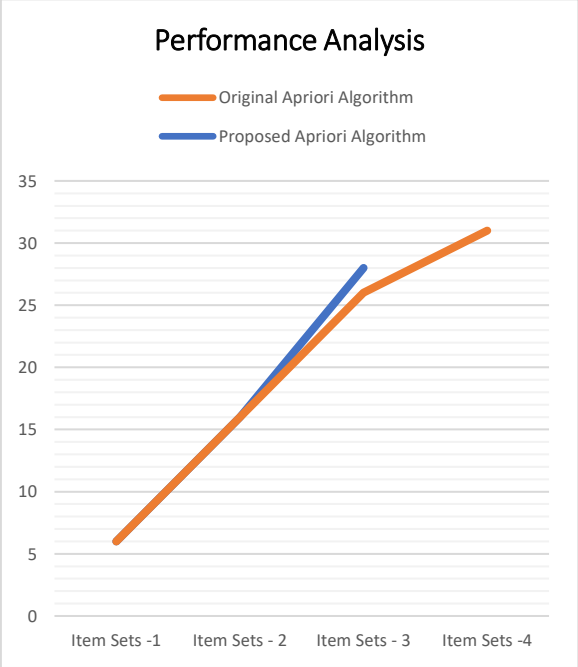

Fig. 3. Performance Analysis of the proposed Apriori algorithm.

From the above graph, it has been observed that, the process for pattern mining for both original and proposed Apriori algorithm similar. However, the proposed Apriori algorithm identifies frequent pattern in the Item Set – 3 itself compared to the original Apriori algorithm. The original Apriori algorithm continuous the process further to identify frequent pattern set which results in inefficient and time consuming. The proposed Apriori algorithm once adapted

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2995 - 3002

will reduces time complexity, input-output load, and usage of memory space. Hence, the performance analysis shows the effectiveness of the proposed Apriori algorithm.

VI. Conclusion

The algorithm presented in this paper is a more refined version of the original Apriori, which has a high input-output load because it scans the database multiple times. It is also slow and requires more storage. Due to a reduction in the number of database scans, the proposed algorithm is significantly faster and more efficient. The proposed Apriori algorithm has reduced time complexity, lesser input-output load and reduction in memory usage.

References

[1] R. Agrawal, T. Imielinski, and A.N Swami, “Mining association rules between sets of items in Large Databases,” Proceedings of the ACM SIGMOD, International Conference on Management of Data, pp.207-216, 1993.

[2] J. S. Park, Ming-Syan Chen and P. S. Yu, "Using a hash-based method with transaction trimming for mining association rules," IEEE Transactions on Knowledge and Data Engineering, vol. 9, no. 5, pp. 813-825, 1997.

[3] J. Han, G. Dong and Y. Yin, "Efficient mining of partial periodic patterns in time series database," Proceedings 15th International Conference on Data Engineering, pp. 106-115, 1999.

[4] Eui-Hong Han, G. Karypis and V. Kumar, "Scalable parallel data mining for association rules," IEEE Transactions on Knowledge and Data Engineering, vol. 12, no. 3, pp. 337-352, 2000.

[5] J. Han, J. Pei and Y. Yin, “Mining frequent patterns without candidate generation,” Sigmod Record, vol. 29, pp. 1-12, 2000.

[6] Wang Tong and HE Pi-lian, “Web log mining by an improved AprioriAll algorithm,” Transaction on Engineering Computing and Technology, vol. 4, pp. 97-100, 2005.

[7] D. Sun, S. Teng, W. Zhang and H. Zhu, "An algorithm to improve the effectiveness of Apriori," 6th IEEE International Conference on Cognitive Informatics, pp. 385-390, 2007.

[8] H. Wang and X. Liu, "The research of improved association rules mining Apriori algorithm", 8th IEEE International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), vol. 2, pp. 961-964, 2011.

[9] Z. Chen, S. Cai, Q. Song and C. Zhu, "An improved Apriori algorithm based on pruning optimization and transaction reduction," 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), pp. 1908-1911, 2011

[10] K. R Suneetha and R. Krishnamoorthi, “Web log mining using improved version of Apriori algorithm,” International Journal of Computer Applications, vol. 29, no. 6, pp. 23-27, 2011.

[11] Y. Jia, G. Xia, H. Fan, Q. Zhang and Xu Li, “An improved Apriori algorithm based on association analysis,”2012 Third International Conference on Networking and Distributed Computing, , pp. 208-211, 2012

[12] J. Singh, H. Ram, Dr. J.S. Sodhi, "Improving efficiency of Apriori algorithm using transaction reduction," International Journal of Scientific and Research Publications, vol. 3, issue. 1, 2013

[13] Ekta Garg, Meenakshi Bansal, “A survey on improved Apriori algorithm,” International Journal of Engineering Research & Technology (IJERT), vol. 2, issue. 7, 2013

[14] S. Dutt, N. Choudhary and D. Singh, “An improved Apriori algorithm based on Matrix data structure,” Global Journal of Computer Science and Technology, vol. 14, issue. 5, version I, 2014

[15] Mohammed Al-Maolegi, Bassam Arkok, “An Improved Apriori Algorithm for Association Rules,” International Journal on Natural Language Computing (IJNLC), vol. 3, no. 1, 2014

[16] P. Yuvraj, Suneetha K. R. , “Modified Apriori graph for frequent pattern mining,” 2016 International Conference on Innovations in information Embedded and Communication Systems (ICIIECS’16), 2016