A Modified Apriori Algorithm for Fast and Accurate Generation of Frequent Item Sets

K.A.Baffour, C.Osei-Bonsu, A.F. Adekoya

Abstract— The Classical Apriori Algorithm (CAA), which is used for finding frequent item sets in Association Rule Mining, consists of two main steps; the join step for generating candidate item sets and the prune step for eliminating candidate item sets that are not frequent.
The CAA despite its simplicity has some limitations; the generation of a large number of candidate item sets, the generation of many combinations that never occur in the database as well as the need to perform several full database scans when generating frequent item sets. In this research, a Modified Apriori Algorithm (MAA) is proposed to address the problem of generating many combinations that never occur in the database by using a row-wise combination generation technique. A comparison of the results of the proposed algorithm against the Classical Apriori Algorithm shows that the proposed algorithm is faster and more efficient. The MAA was implemented on transaction databases and the results were compared against results from four (4) other Improved Apriori Algorithms for efficiency. The results of the comparative analysis showed that the MAA was more efficient in terms of execution time than the other Improved Apriori Algorithm.

Index Terms: Association Rule Mining, Candidate item sets, Combinations, Database, Frequent item sets, Join, Prune.

1 Introduction

The world has gradually evolved into a knowledge economy which according to Powell and Snellman (2004) is an economy in which production of goods and services is based on knowledge-intensive activities that contribute to an accelerated pace of technical and scientific advancement. Over the years so much data mostly in the form of files and databases have been generated by many organizations.

Data Mining

Data mining is the process of collecting and aggregating data from different perspectives and sources and analyzing the data to generate meaningful information.

Association Rule Mining

Association rule mining, introduced in 1993, is one of the most useful applications of data mining. Association rule mining makes it possible to discover patterns and interesting relationships between items in databases (Wakchaware, 2014).

Classical Apriori Algorithm

The Classical Apriori Algorithm (CAA), which is used for finding frequent item sets, was developed by Aggrawal and Srikant in 1994. The CAA is very simple to implement and consists of two main steps; the join step for generating candidate item sets and the prune step for eliminating candidate item sets that are not frequent (Kaur, 2014).
The CAA is the traditional algorithm used to generate Association Rules. The CAA despite its simplicity and ease of implementation has several drawbacks. Some weaknesses of the algorithm include:

- Generation of many candidate item sets consisting of many infrequent and unnecessary item sets
- the generation of a large number of combinations that never occur in the database as well as
- the need to perform several full database scans when generating frequent item sets .

1.1 Aims

In this paper, a Modified Apriori Algorithm (MAA) is proposed to solve a major problem of the CAA where many combinations that do not exist in the database are generated. The MAA addresses this issue by using the principle of generating combinations from the frequent items found in each row of the transaction database.

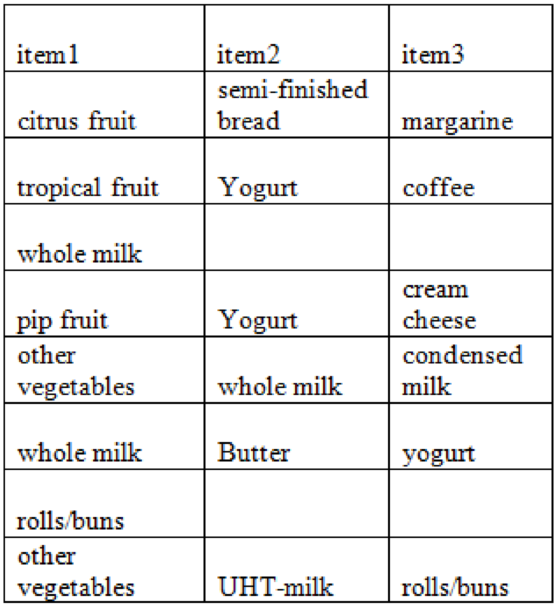
2 RESEARCH METHODOLOGY

The MAA and its architecture are explained in this section.

2.1 Data Collection

Secondary data, "Groceries Dataset" obtained from 'http: //www.inside-r.org/packages/cran/arules/docs/Groceries' was used. A transactions database, transactions_db was created on a database server (MySQL).Inside this database, table, Groceries was created and populated with the content of the Groceries Dataset.

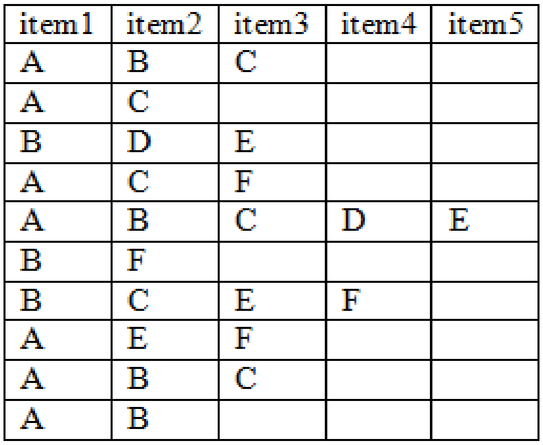
TABLE 1
SECTION OF GROCERIES DATASET



item1	item2	item3
citrus fruit	semi-finished bread	margarine
tropical fruit	Yogurt	coffee
whole milk		
pip fruit	Yogurt	cream cheese
other vegetables	whole milk	condensed milk
whole milk	Butter	yogurt
rolls/buns		
other vegetables	UHT-milk	rolls/buns

TABLE 2

SAMPLE DATASET FOR ILLUSTRATION



item1	item2	item3	item4	item5
A	B	C		
A	C			
B	D	E		
A	C	F		
A	B	C	D	E
B	F			
B	C	E	F	
A	E	F		
A	B	C		
A	B			

2.2 Data Preprocessing

Data preprocessing involves transforming raw data into an understandable format. The data processing involves transformation of raw data into appropriate format. Using Microsoft Excel (Spreadsheet Application), the following three (3) activities of Data Processing were carried out on the Groceries dataset:

- replacement of all forward slash '/' with the word 'or',
- replacement of all open bracket ' (' with the word 'and', replacement of closing bracket ')' with a blank space and
- replacement of all hyphen '-' with a blank space.

2.3 Architecture of the MAA

The MAA consists of six (6) modules namely;

- Database connection and table selection module
- Minimum Support Threshold Definition module
- Frequent Items module
- Array of frequent items module
- Combination generation module
- Frequent item sets module

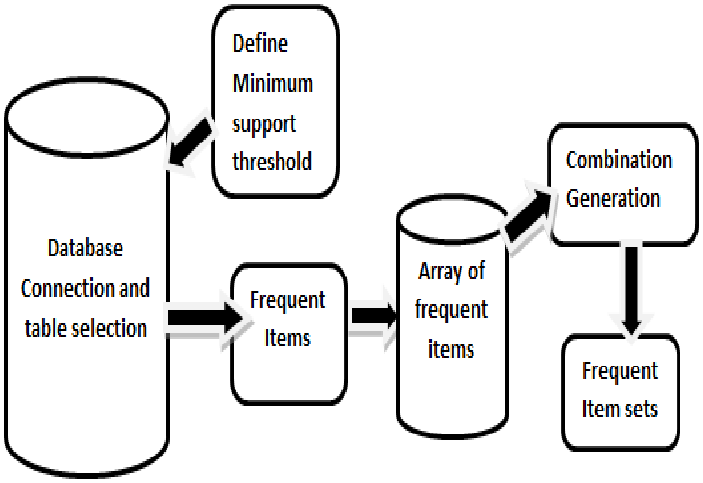


Fig. 1. Architecture of MAA

2.4 Database Connection and Table Selection

All databases reside on a database server. The database server used for this research was the MySQL database server. To access the transactions_db database, a connection had to be established with the database server. After connecting to the database server, the Groceries table was selected.

2.5 Minimum Support Threshold (MST) Definition

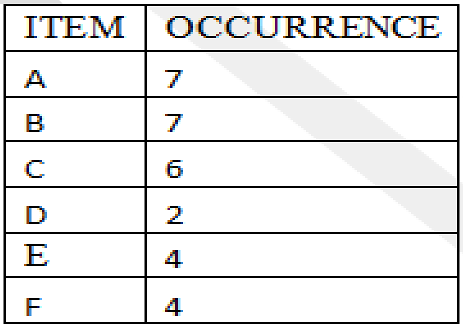
The MST is used in discovering frequent items and item-sets. The MST value is usually accepted as an input from the user since it is user defined.

For this research however, the MST value was determined using Measures of Central Tendency. Measures of Central Tendency are statistical measures that make it possible to choose a value that best describes or represents a whole set of data. The MST for this research was determined using the measure known as the Mid-Range. The Mid-Range of a set of data is the mean of the maximum and minimum values in the dataset. The Mid-Range of the occurrence counts of the items in the transactions database was calculated and used as the MST. It is defined by:

$$M = (max + min) / 2 \qquad (1)$$

TABLE 3

OCCURRENCE COUNT OF ITEMS



ITEM	OCCURRENCE
A	7
B	7
C	6
D	2
E	4
F	4

From Table 3 Max occurrence is 7 and the Minimum occurrence is 2.

$$M = (7+2) / 2 = 4.5 \approx 5$$

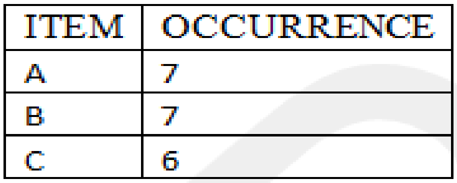
Therefore the MST is 5.

2.6 Frequent Items

Frequent items refer to those items with occurrence counts greater than or equal to the MST.

TABLE 4

FREQUENT ITEMS

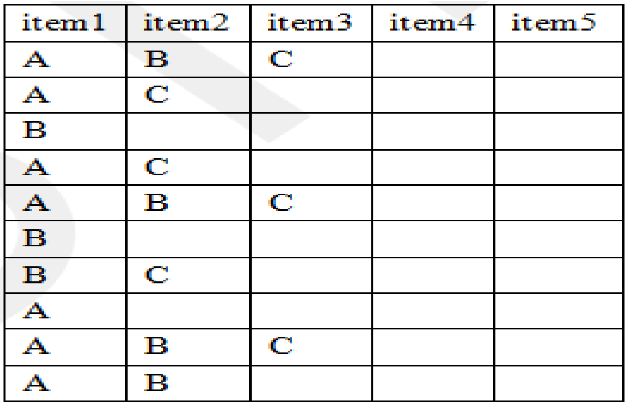


ITEM	OCCURRENCE
A	7
B	7
C	6

2.7 Array of Frequent Items

After the generation of frequent items, an empty array with the same structure as the Groceries table was created. The contents of the Groceries table were read into the array to populate it. The infrequent items were then deleted from the array.

TABLE 5
STRUCTURE OF ARRAY OF FREQUENT ITEMS



item1	item2	item3	item4	item5
A	B	C		
A	C			
B				
A	C			
A	B	C		
B				
B	C			
A				
A	B	C		
A	B			

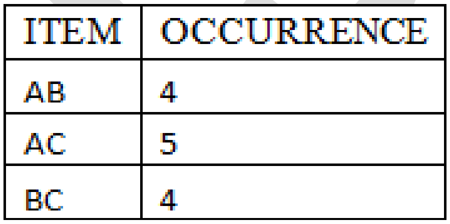
2.8 Combination Generation

Starting with the first row to the last row in the array of frequent items, all possible combinations were generated using the items from that row.
Row1: AB, AC, BC
Row2: AC
Row3: -
Row4: AC
Row5: AB, AC, BC
Row6: -
Row7: BC
Row8: -
Row9: AB, AC, BC
Row10: AB

2.9 Frequent Item sets

Frequent item sets were obtained by counting the occurrence of each unique combination and checking if it was greater than or equal to the MST. All those combinations that met this criterion were selected as frequent item sets.

TABLE 6
ITEMSET OCCURRENCE



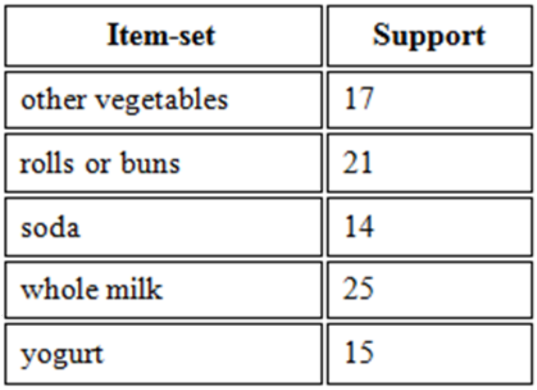
ITEM	OCCURRENCE
AB	4
AC	5
BC	4

From the Table above, the only combination that meets the MST is AC.

3 RESULTS

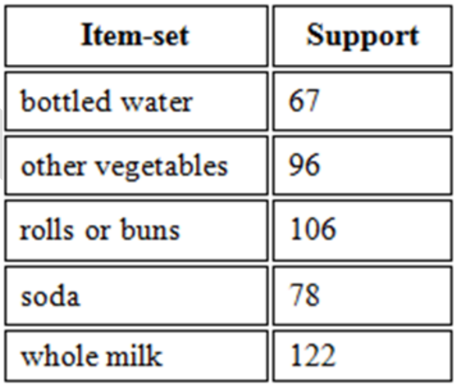
The results of implementing the MAA on the extracts of the Groceries dataset are presented in this section. The comparative analyses of the MAA against the other Improved Apriori Algorithms are also presented.

TABLE 7
RESULTS (MAA ON FIRST 100 RECORDS)



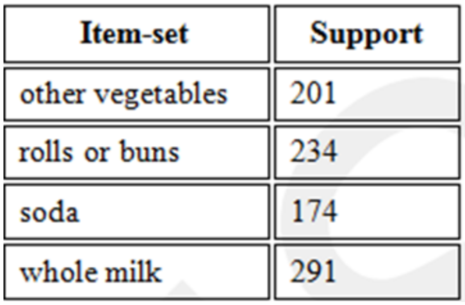
Item-set	Support
other vegetables	17
rolls or buns	21
soda	14
whole milk	25
yogurt	15

TABLE 8
RESULTS (MAA ON FIRST 500 RECORDS)



Item-set	Support
bottled water	67
other vegetables	96
rolls or buns	106
soda	78
whole milk	122

TABLE 9
RESULTS (MAA ON FIRST 1000 RECORDS)



Item-set	Support
other vegetables	201
rolls or buns	234
soda	174
whole milk	291

3.1 Comparative Analysis of MAA against Other Improved Apriori Algorithms (IAA)

The MAA was comparatively analyzed against three (3) other Improved Apriori Algorithms for efficiency.

3.2 MAA against Matrix Apriori

The MAA was compared against the Matrix Apriori Algorithm proposed by Luo and Wang (2010).

TABLE 10
EXECUTION TIMES FROM THE MAA AND MATRIX APRIORI ALGORITHM

Database size	Matrix Apriori(seconds)	MAA(seconds)
500	35	0.138
1000	108	0.044
1500	178	0.065
2000	214	0.123
2500	282	2.182

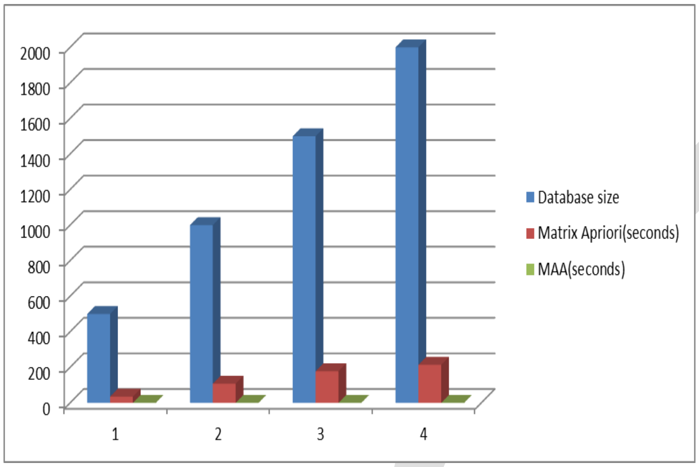


Fig. 2. Column Chart comparing the execution times of the MAA and Matrix Apriori Algorithm

3.3 MAA against Record Filter Apriori Algorithm

The MAA was compared against the Record Filter Apriori Algorithm proposed by Anshu and Raghuvashi (2010).

TABLE 11
EXECUTION TIMES FROM THE MAA AND RECORD FILTER APRIORI ALGORITHM

Database size	Record Filter(seconds)	MAA(seconds)
500	37	0.138
1000	82	0.386
1500	149	0.406
2000	348	1.643

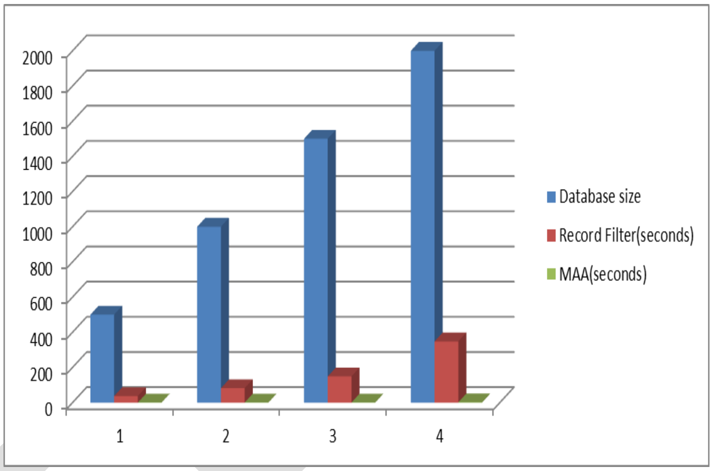


Fig. 3. Column Chart comparing the execution times of the MAA and Record Filter Apriori Algorithm

3.4 MAA against Frequent Item set Prediction Apriori Algorithm

The MAA was compared against the Frequent Item set Prediction Apriori Algorithm proposed by Ayman and Alsheref (2014) using an MST of three (3).

TABLE 12
RESULTS FROM THE MAA AND THE FREQUENT ITEM SET PREDICTION APRIORI ALGORITHM

MAA		FREQUENT ITEMSET PREDICTION	
Item set	Support	Item set	Support
I1	6	I1	6
I2	7	I2	7
I2, I1	4	I3	5
I3	5	I4	3
I3, I1	4	I1,I2	4
I3, I2	3	I1,13	4
I4	3	I2,I3	3
I4, I2	3	I2,I4	3

3.5 Discussion and Analysis of Results

From the Tables 10 to 12 and Figures 2 to 3 depicting the results of the comparative analysis of the proposed MAA against other IAA, it can be succinctly shown that the MAA generates exactly the same number and size of item sets as the other IAA i.e., the MAA generates the same results as the other IAA. With regards to execution time however, the MAA requires very little time for generating its results and can therefore be said to be more efficient and computationally less expensive than the other IAA.

4 CONCLUSION

In this research, an enhanced Modified Apriori Algorithm (MAA) was developed to address a major problem of the CAA From the results, it was observed that the MAA was better and more efficient than the CAA since it succeeded in eliminating the problem of non-existent combination generation.

5 REFERENCES

[1] Aggarwal, S., & Sindhu, R. (2015). An approach to improve the efficiency of apriori algorithm (No. e1410). PeerJ PrePrints.

[2] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

[3] Anshu, C., & Raghuvanshi, C. S. (2010). An algorithm for frequent pattern mining based on Apriori. International Journal on Computer Science and Engineering, 2(04), 942-947.

[4] Ayman, K.E., Alsheref, F.K. (2014). Improved Apriori Algorithm VIA Frequent Item sets Prediction. International Journal of Advanced Research in Computer Science and Engineering, 4(4), 748-758.

[5] Dutt, S., Choudhary, N., & Singh, D. (2014). An Improved Apriori Algorithm based on Matrix Data Structure. Global Journal of Computer Science and Technology, 14(5), 1-5.

[6] Garg, S., & Sharma, A. K. (2013). Comparative Analysis of Various Data Mining Techniques on Educational Datasets. International Journal of Computer Applications, 74(5), 1-5

[7] Kaur, G. (2014). Improving the Efficiency of Apriori Algorithm In Data Mining. International Journal of Science, Engineering and Technology, 2(5), 315-326.

[8] Luo, X., & Wang, W. (2010, June). Improved algorithms research for association rule based on matrix. In Intelligent Computing and Cognitive Informatics (ICICCI), 2010 International Conference on (pp. 415-419). IEEE (Conference proceedings)

[9] Powell, W. W., & Snellman, K. (2004). The knowledge economy. Annu. Rev. Sociol., 30, 199-220

[10] Wakchaware, S. (2014). Large Databases–Association Rule Mining. International Journal for Research in Emerging Science and Technology, 1(4), 19-26.