Scientific Journal of Impact Factor (SJIF): 4.14
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 6, June -2016

To Enchanced & Optimize The Apriori Algorithm Using Tokenization Based Association Rule Mining

Komal Thakur[1], Vinay Chopra[2]

[12]*Department of Computer Science and Engineering, DAVIET Jalandhar*

Abstract — *In the term of Data Mining, Association Rule mining has remarkable role. Association rule mining is a popular mining technique that identifies interesting correlations between database attributes. This research paper gives the detailed introduction to proposed tokenization approach based on the apriori algorithm of association rule mining.*

Keywords- *association rules, frequent item sets, tokens.*

I. INTRODUCTION

A. Data Mining- Data mining is a process that discovers the knowledge or hidden patterns from the large databases. Data mining is also known as the core processes of Knowledge Discovery in Databases (KDD). The KDD process is commonly defined as the following stages:

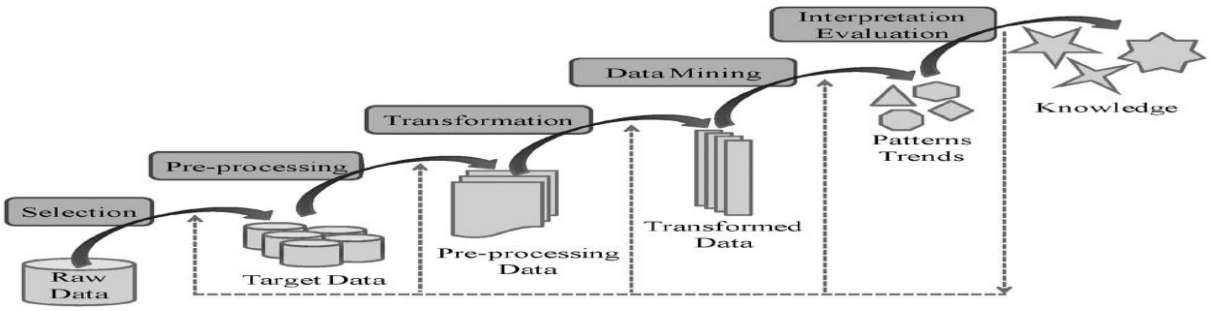

Figure 1. KDD Process

Data mining involves six main classes of tasks:

1. Anomaly detection - The identification of unusual records or data errors that require further investigation.

2. Association rule learning (Dependency modeling) - This searches the relationships between the variables.

3. Clustering- The clustering technique defines the classes and put objects in each class. Clustering is a DM technique which makes useful cluster of objects having similar characteristics using automatic technique.

4. Classification- This is used to classify each item in a set of data into one of pre defined set of classes or groups. Classification method uses mathematical techniques such as decision trees, linear programming, and neural network.

5. Regression- This attempts to find a function which models the data with the least errors.

6. Summarization- This provides a more compact representation of the dataset including visualization and report generation.

B. Association Rule Mining- Association Rule Mining are the data mining function that just discovers the probability of co-occurrence of items in a collection or dataset. As in the example, here are the 5 transactions in which in transaction has some items and in the next table, there is the probability of occurrence items i.e. in the T1 transaction there are only two items- (Bread, Milk). The (Bread, Milk) will be written as 1 and all remaining will written as 0.

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

	Beer	Bread	Milk	Diaper	Eggs	Coke
T_1	0	1	1	0	0	0
T_2	1	1	0	1	1	0
T_3	1	0	1	1	0	1
T_4	1	1	1	1	0	0
T_5	0	1	1	1	0	1

Figure 2. Association Rule Mining

C. ALGORITHM USED FOR ARM- The two important Algorithm used for association rule mining are apriori algorithm and FP-Tree algorithm. The apriori algorithm is used for finding patterns called frequent item sets. A frequent item set is a set of items appearing together in a number of database records meeting a user specified threshold. The FP- Tree algorithm is to partition the original database to smaller sub-databases by some partition cells, and then to mine item sets in these sub databases. The FP-Tree construction takes exactly two scans of the transaction database. The first scan collects the set of frequent items , and second scan constructs the FP- Tree.

II. LITERATURE REVIEW

[1]Sotiris Kotsiantis, DimitrisKanellopoulos,*"Association Rules Mining: A Recent Overview"* GESTS International Transactions on Computer Science and Engineering, vol.32 (1), pp 71-82, 2006.
In this paper, they provide the preliminaries of basic concepts about association rule mining and survey the list of existing association rule mining techniques.They also describes the methods that hadproposed for increasing the efficiency of association rules algorithms.

[2]Rakesh Kumar Soni1, Neetesh Gupta, Amit Sinhal, "*An FP-Growth Approach to MiningAssociation Rules*" International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 2, February 2013, pp 1 – 5.
The authors improved theperformance of mining in this paper. They use Sampling Technique to convert text document in to the appropriate format. This format containsdata in the form of word and topic of word. This format take as a input in FP-Growth algorithm for givensupport value and get association rules of that transaction data, and after getting association rules applyclustering process and then get clusters for that association rules.

[3]JaiWeiHan, Jian Pei, Yiwen Yin &Runying Mao, "*Mining frequent patterns without candidate generation: A Frequent pattern tree approach*" Data mining and knowledge discovery, Netherlands, pp 53-87, 2004.
In this paper, they proposed a novel frequent-pattern tree (FP-tree) structure, which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns, and develop an efficient FP-treebased mining method. Their performance study shows that the *FP-growth* method is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the *Apriori*algorithm and also faster than some recently reported new frequent-pattern mining methods.

[4] Huan Wu, Zhigang Lu, Lin Pan, RongSeng XU and Wenbaojiang "*An improved Apriori based algorithm for association rule mining*" IEEE Sixth international conference on fuzzy systems and knowledge discovery, pp 51-55, 2009.
In this paper, based on theoriginal Apriori algorithm, an improved algorithm IAA was proposed by the authors. IAA adopts a new count-based method to prunecandidate itemsets and uses generation record to reduce totaldata scan amount.

[5] Badri Patel, Vijay K Chaudhari, Rajneesh K Karan, YK Rana "*Optimization of Association Rule Mining Apriori Algorithm using ACO*" International Journal of Soft Computing and Engineering vol 1, issue 1, pp 24-26, March 2011.
Onthe basis of the association rule mining and Apriori algorithm,the authors proposed an improved algorithm based on the AntColony Optimization algorithm. They optimize the resultgenerated by Apriori algorithm using Ant colony optimizationalgorithm. The algorithm improved result produces by Apriorialgorithm.

[6] K.Saravana Kumar, R.ManickaChezian,"*A Survey on Association Rule Mining using Apriori Algorithm*" International Journal of Computer Application, vol. 45, no. 5, pp 47-50, May 2012.
In this paper, the author surveys the most recent existing association rule mining techniques using Apriori algorithm. The conventional algorithm of association rules discovery proceeds in two steps. All frequent item sets are found in the first step. The frequent item set is the item set that is included in at least minimum support transactions. The association rules with the confidence at least minimum confident are generated in the second step.

[7] Rafael S. Parpinelli, Heitor S. Lopes, Alex A. Freitas, *"Data Mining With an Ant Colony Optimization Algorithm"* IEEE Transactions on evolutionary computing, vol. 6, no. 4, pp 321-332, August 2002
The authors proposed an algorithm for data miningcalled Ant-Miner (ant-colony-based data miner). The goal of AntMiner is to extract classification rules from data.theycompare theperformance of Ant-Miner with CN2, a well-known data miningalgorithm for classification, in six public domain data sets. The results provide evidence that:

I. Ant-Miner is competitive with CN2with respect to predictive accuracy
II. The rule lists discoveredby Ant-Miner are considerably simpler (smaller) than those discovered by CN2.

[8] SuhaniNagpal "Improved *Apriori Algorithm using logarithmic decoding and pruning*" International Journal of Engineering Research and Applications, vol. 2, issue 3, pp. 2569-2572, May-Jun 2012.

In this paper, the authorimproves the performance of the conventionalApriori algorithm that mines the association rules.The approach is to attain the desired improvement is to create a more efficient new algorithm out of theconventional one by adding the encoding anddecoding mechanisms to the latter in order todemonstrate the importance of the efficient decodingto high data mining performance and from variousexperiments it is proved that the logarithmicdecoding method is the most efficient among the allmethods it can speed up all the required processes.

[9] Fernando E. B. Otero, Alex A. Fretas and Colin G. Johnson "*A new sequential covering strategy for inducing classification rules with ant colony algorithms*" IEEE transaction on evolutionary computation, vol. 17, no. 1, pp 64-76, February 2013.
The author proposes a new sequential covering strategyfor ACO classification algorithms to mitigate the problem of ruleinteraction, where the order of the rules is implicitly encodedas pheromone values and the search is guided by the qualityof a candidate list of rules.

[10] Sang Jun Lee, KengSiau*"A review of data mining techniques"* Industrial Management and Data Systems, University of Nebraska-Lincoln Press,USA,pp41-46,2001
In this paper, the author discussed the requirements and challenges of data mining. Theauthor also describes about the major data mining techniques such as statistics, artificial intelligence, decision tree approach, genetic algorithms and visualization.

[11] M. Dorigo, V. Maniezzo, A. Colorni, "*The ant system: optimization by a colony of cooperating agents",* IEEE Trans. Systems Man Cybernet. B 26, pp 29–42, 1996.

The author proposed the methodology to the classical TravelingSalesman Problem (TSP), and report simulation results. They alsodiscuss parameter selection and the early setups of the model,and compare it with tabu search and simulated annealing usingTSP. And also discussed thesalient characteristics-global data structure revision, distributedcommunication and probabilistic transitions of the AS.

[12] MeeraNarvekar, ShafaqueFatma Syed, "*An optimized algorithm for association rule mining using FP tree",*International Conference on Advanced Computing Technologies and Application, pp 101-110,2015

In this paper, the author designed a new technique which mines out all the frequent item sets without the generation of the conditional FP trees. Unlike FP tree it scans the database only once which reduces the time efficiency of the algorithm. It also finds out the frequency of the frequent item sets to find out the desired association rules.

III. PROPOSED METHODOLOGY

Proposed technique uses tokens i.e. those item-sets that co-occurrence with representative item can be identified quickly and directly using this simple and quickest token based method. This will avoid redundant operations of item-sets generation and many frequent items having the same supports as representative item, so the cost of support count is reduced hence the efficiency is improved. The proposed algorithm uses the following steps:
1. Scanning the database and converting it into vertical data format.
2. Generating Trans_tokenSet from vertical data format and also maintaining a list for no of iterations.
3. Finding the Frequent 1 itemset from the Trans_tokenSeti.e. the length of Trans_tokenSet of the item sets.
4. Sorting the itemset according to the ascending order of the Trans_tokenSet by its minimum support.
5. Gather these items as Keysets from the Trans_tokenSet.
6. Generate the Bit Table for each keythat is available in Items keyset
7. Generate Subsume for each item in Items keyset
8. for each item in Items

```
If item. Subsume<> " "
If item. Support == min_sup then
FindItemsetsEqualsMinSup(item, item. Support)
 Else
FindItemsetsGreaterThanMinSup (item, item. Support)
 End If
Else
If item. Support>min_sup
      AND Item_Sequence<Item.Length Then
         FindItemsetSubnumeNone(item)
Endif
Endif
```


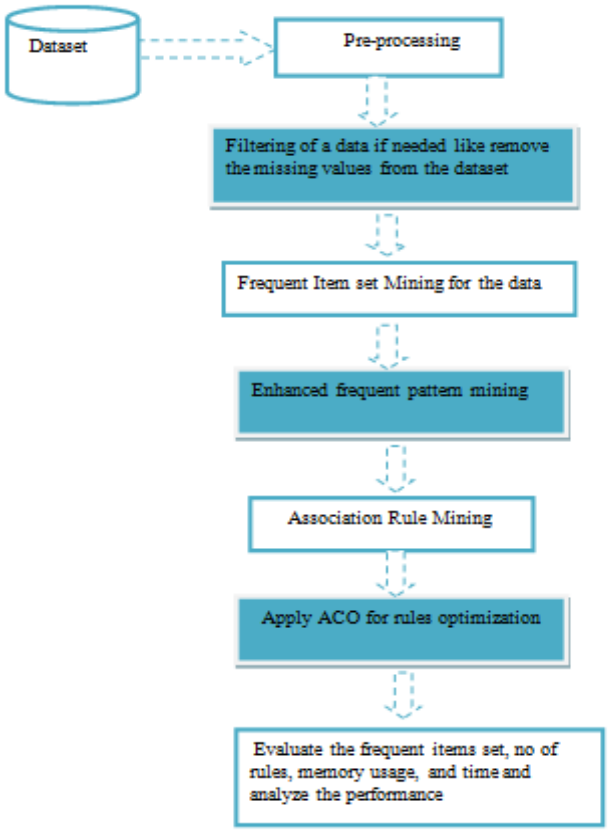

Figure 3. Flowchart of proposed algorithm

IV. EXPERIMENTAL SETUP

This algorithm is implemented in the Java Language because it contains the data set mining to develop the applications in the data mining and tools. The snapshot of frequent item sets mining is as follows which are acquired during implementation.

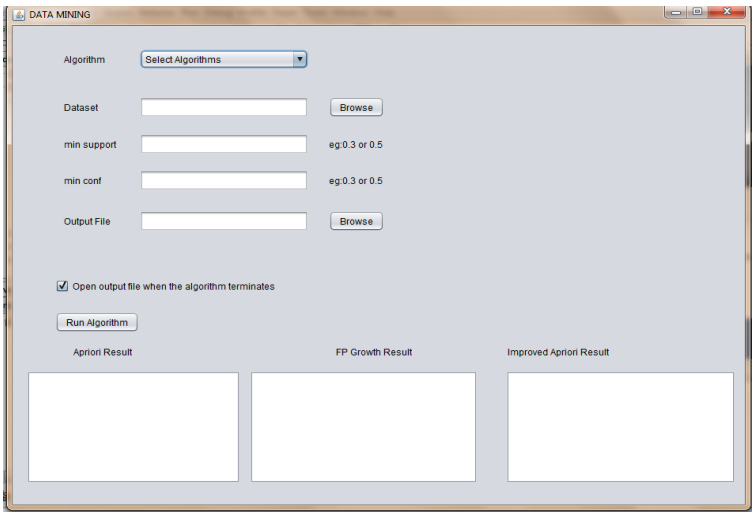

Figure 4. The screenshot for selecting the algorithm and the parameters for finding the minimum support and confidence for improved Apriori result

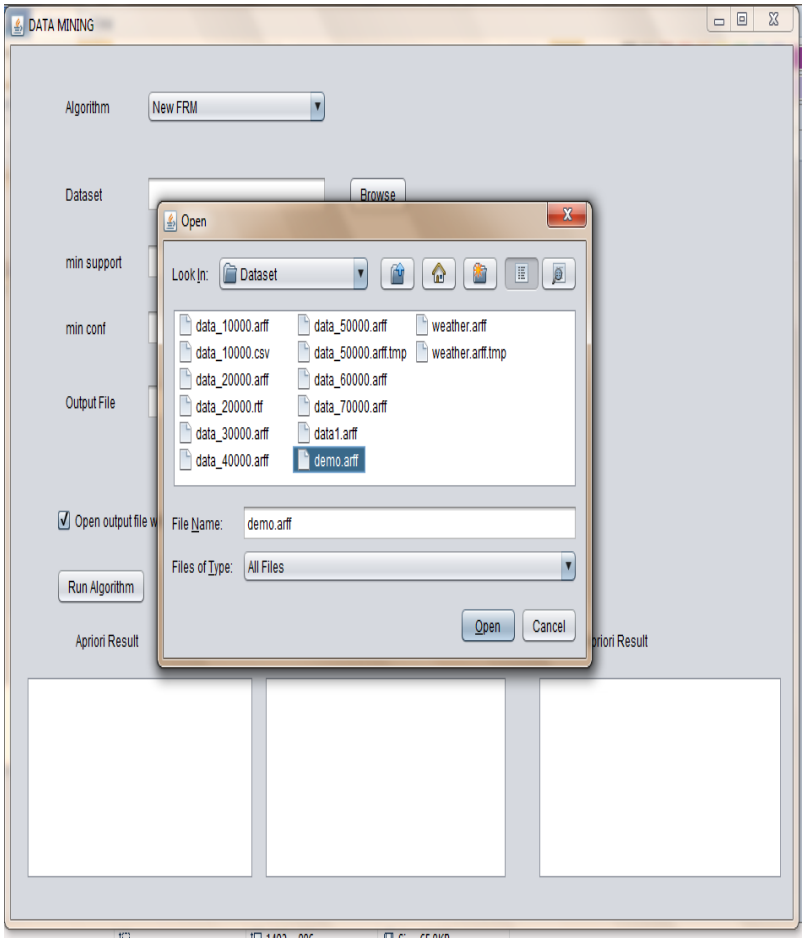

Figure 5. The screenshot for choosing the dataset.

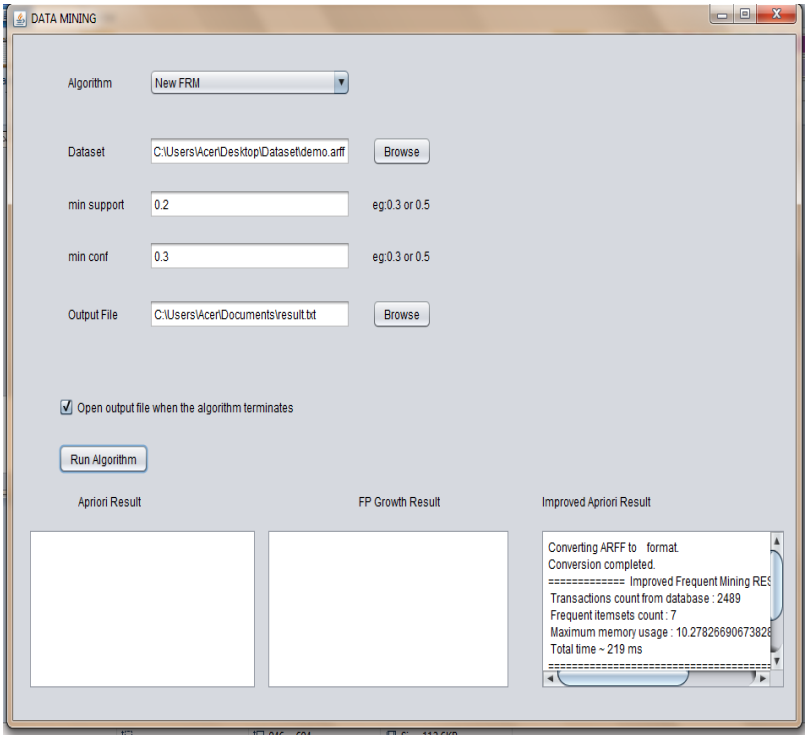

Figure 6. The screenshot for the final output describing the parameters such as memory usage, time consumed and frequent itemset count are generated.

V. REFERENCES

[1] Sotiris Kotsiantis, DimitrisKanellopoulos, *"Association Rules Mining: A Recent Overview"* GESTS International Transactions on Computer Science and Engineering, vol.32 (1), pp 71-82, 2006.

[2] Rakesh Kumar Soni1, Neetesh Gupta, Amit Sinhal, "*An FP-Growth Approach to MiningAssociation Rules"* International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 2, February 2013, pp 1 – 5.

[3] JaiWeiHan, Jian Pei, Yiwen Yin &Runying Mao, "*Mining frequent patterns without candidate generation: A Frequent pattern tree approach*" Data mining and knowledge discovery, Netherlands, pp 53-87, 2004.

[4] Huan Wu, Zhigang Lu, Lin Pan, RongSeng XU and Wenbaojiang "*An improved Apriori based algorithm for association rule mining*" IEEE Sixth international conference on fuzzy systems and knowledge discovery, pp 51-55, 2009.

[5] Badri Patel, Vijay K Chaudhari, Rajneesh K Karan, YK Rana "*Optimization of Association Rule Mining Apriori Algorithm using ACO*" International Journal of Soft Computing and Engineering vol 1, issue 1, pp 24-26, March 2011.

[6] K.Saravana Kumar, R.ManickaChezian,"*A Survey on Association Rule Mining using Apriori Algorithm*" International Journal of Computer Application, vol. 45, no. 5, pp 47-50, May 2012.

[7] Rafael S. Parpinelli, Heitor S. Lopes, Alex A. Freitas, *"Data Mining With an Ant Colony Optimization Algorithm"* IEEE Transactions on evolutionary computing, vol. 6, no. 4, pp 321-332, August 2002

[8] SuhaniNagpal "Improved *Apriori Algorithm using logarithmic decoding and pruning*" International Journal of Engineering Research and Applications, vol. 2, issue 3, pp. 2569-2572, May-Jun 2012.

[9] Fernando E. B. Otero, Alex A. Fretas and Colin G. Johnson "*A new sequential covering strategy for inducing classification rules with ant colony algorithms*" IEEE transaction on evolutionary computation, vol. 17, no. 1, pp 64-76, February 2013.

[10] Sang Jun Lee, KengSiau*"A review of data mining techniques"* Industrial Management and Data Systems, University of Nebraska-Lincoln Press, USA, pp41-46, 2001

[11] M. Dorigo, V. Maniezzo, A. Colorni, "*The ant system: optimization by a colony of cooperating agents",* IEEE Trans. Systems Man Cybernet. B 26, pp29–42, 1996

[12] MeeraNarvekar, ShafaqueFatma Syed, "*An optimized algorithm for association rule mining using FP tree",* International Conference on Advanced Computing Technologies and Application, pp 101-110,2015

