Radical Analysis Review on Improved Versions of Apriori Algorithm for Association Rule Mining

Abhijit Sarkar[1], Biswajit Jana[2], Abhirup Paria[3], Tarun Kumar Ghosh[4], Shaon Bandyopadhyay[5]

[1,2,3,4]: Dept. of CSE(Specialization), Haldia Institute of Technology, W.B.,India.

[5]:Dept of CSE, Haldia Institute of Technology, W.B., India.

ABSTRACT:

Worthwhile information can be effectively abridged by the practice of Data mining. ARM namely Association rule mining, considered to be a unique significant data mining utility for locating common as well as frequent itemset. Apriori tactic in data mining is one of the most widely utilized strategies for discerning the frequent itemset in a transaction database. It extracts intriguing connections, common arrangements as well as linkages amid sets of objects from transaction repositories or databanks sources. Mainly the available techniques demand numerous runs through the database for discovering common patterns leading to a substantial number of record reads and exerting a massive strain on the input/output subsystem. Efficiency has been engaged for many times in data processing for detecting common itemsets. Large series of reforms has been presented in the similar research to mitigate the insufficiencies of Apriori algorithm.

KEYWORDS*: Apriori, ARM, confidence, frequent itemset.*

I. INTRODUCTION

Investigating vast pre-existing databases in generating new patterns and insights is the progression of data mining. Data Mining can be defined as wisdom finding. It is used to identify patterns and correlations from the vast database. Various protocols have stayed proposed in data mining wherein, ARM is vital for locating frequent patterns. Apriori is among the regularly used strategies for the association rule mining. The Apriori method produce recurring traits from repository whose support should fulfill the predefined threshold requirement where these relevant results are used to build association rule where confidence must achieve the minimal expectation criteria.

The main purpose of our work is to survey the various versions of optimized apriori algorithm. In all previous survey papers, only the description about some improved algorithm were given. Whereas, in our survey paper we are providing the summary of algorithm theory, advantages and disadvantages, as well as an example. As all examples are calculated using same sample transaction database, we can easily understand the differences in different algorithm or difference in frequent itemset generation process. We can also get the information about which algorithm produce more candidate itemsets than other.

II. APRIORI ALGORITHM

TID	List of items
T1	I1, I2, I3, I5
T2	I2, I3, I4
T3	I1, I3, I4
T4	I1, I2, I5
T5	I1, I2, I3, I5
T6	I1, I5
T7	I1, I4, I5
T8	I2, I3, I4

Figure 1: Database containing transaction info.

Figure 1 illustrates the transaction database. It comprises eight individual transactions and the things bought inside the trade.

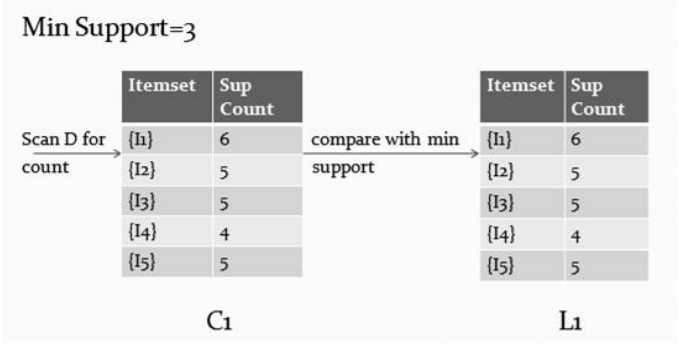

Figure 2: Synthesis of probable itemset and frequent 1-itemset

Figure 2 illustrates the minimum support of distinct 1-itemsets as well as the chosen frequent 1-itemsets based on weighted support predefined threshold 3 after trimming.

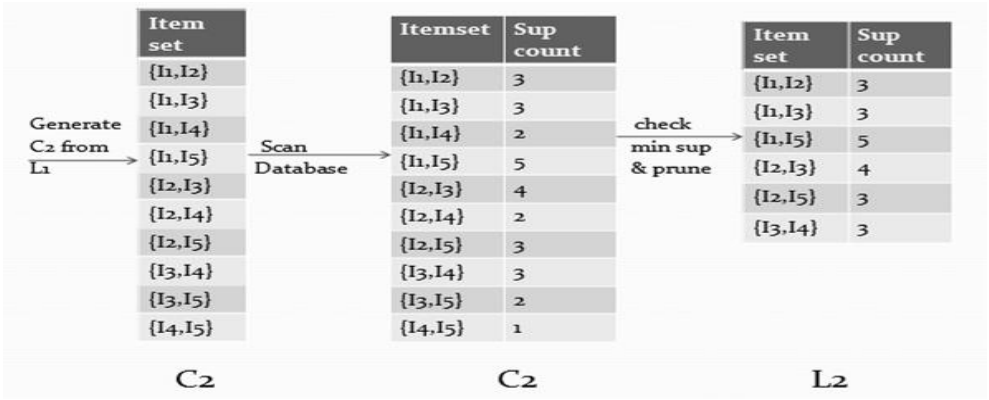

Figure 3: Synthesis of candidate itemset as well as successive frequent 2-itemset

Above picture displays the support count of specific candidate 2-itemsets and the identification of frequent 2-itemset based on the required support target value 3.

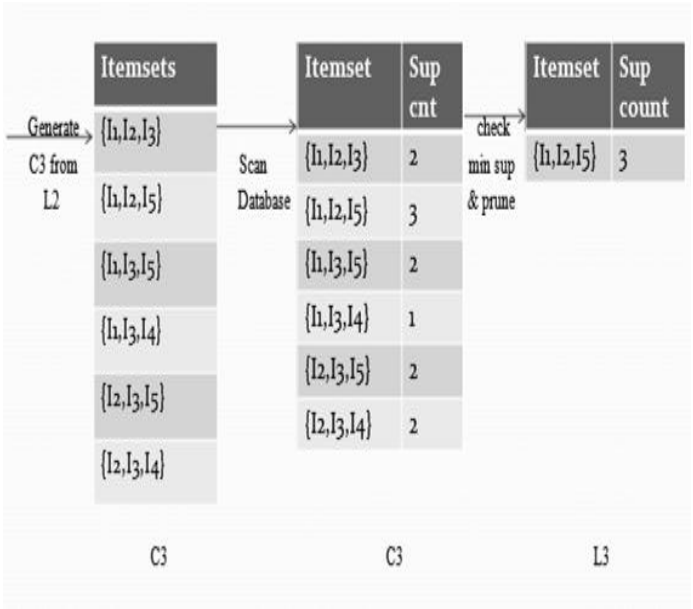

Figure 4: Synthesis of candidate 3-itemset as well as successive frequent 3-itemset

Figure 4 illustrates the support count of several candidate 3-itemsets as well as the determined frequent 3-itemset depending on minimal support target value 3 after trimming and cessation of frequent itemset creation. The association rule set which would be the concluding frequent itemset comprises {I1,I2,I5}.

Association rules	Confidence	Status
R1: I1^I2-> I5	Sc{I1,I2,I5}/Sc{I1^I2}=3/3=100%	Selected
R2: I2^I5->I1	Sc{I1,I2,I5}/Sc{I2,I5}=3/3=100%	Selected
R3: I1^I5->I2	Sc{I1,I2,I5}/Sc{I1,I5}=3/5=60%	Rejected
R4: I1->I2^I5	Sc{I1,I2,I5}/Sc{I1}=3/6=50%	Rejected
R5: I2->I1^I5	Sc{I1,I2,I5}/Sc{I2}=3/5=60%	Rejected
R6: I5->I1^I2	Sc{I1,I2,I5}/Sc{I5}=3/5=60%	Rejected

Figure 5: Enumerate of chosen association rules based upon confidence criterion = 70 percent

Figure 5 displays the numerous possible association rules from the item occurring in association rule set plus determination of association rules R1 and R2 amongst them relying on minimal confidence cutoff = 70 percent.

III. RELATED WORKS

This section includes literatures carried out on Improvement of Apriori Algorithm. Following are the improvement proposed over Apriori algorithm:

In [17] writers enhanced this method significantly. The algorithm runs as follows - choose a randomly selected S from provided data, and then seek for frequent items rather than provided data. So, it employs a lower bit criterion than minimal support count to locate the common itemsets near to itemset which is considered to be frequent.

In [13] researchers developed the hash-based concept assisted DHP algorithm. The candidates created utilizing this approach is substantially minorcompared to prior methods, notably obtained candidate 2-itemsets. Production of reduced candidate sets drastically lower the transaction dimensionality of the data at a much previous assertion of the rounds.

In [15] authors proposed the partition-based method. This approach searches the database just twice, therefore lowering the I/O Activity considerably and hence improving the effectiveness of the process. High capacity datasets are conceptually separated into numerous disjunct chunks that are used to build locally frequent item sets, and these item sets are utilised to predict the desired global similar patterns sets by verifying its support.

In [8] authors provided filtering strategy rather than trimming, in which often and uncommon itemsets are formed. They further refined the screening approach [9] by examining just (k-1) uncommon itemset as separators.

In [5] Authors introduced Upgraded DC Apriori method, which modified the retention database structure, strengthened connectivity of frequent item sets, improved joining step, and considerably decreased the number of links. Authors suggested aninnovative technique for the very purpose of enhancing the speed of Apriori, termed BitApriori [6]. The technique mines frequent itemsets. In BitApriori, the information model twofold string is designed to database descriptor. The count of support may be accomplished by using the Bitwise “And” function over the so called bit numbers.

Added strategy provided in this study is a particular comparable trimming. The Expanded BitApriori algorithm [20] is identical to a BitApriori algorithm, but the key distinction is that in BitApriori were using the Bitwise "While" activity on binary representation and in Expanded BitApriori Bitwise "XOR" action has been used on binary representation. Ranking is not dependent on diminishing support count [20].

In [2] authors utilised adjacency matrix as optimal data model. These approach lowers the volume of candidate-K itemset in each subsequent repetition. Pruning is however performed at two phases which lessens the storage space. The algorithm works as follow: **(**Frequent item set generation:) First select max item from sorted table as frequent Itemset name given as A. Items of adjacency matrix which are associated with A are considered as candidate itemsets. This procedure is continued until no more itemsets are generated.

In [7] the data mapping technique is optimized, so the database is iterated only once. After generating frequent 1-itemset following frequent item sets can be get by scanning Ck and Lk. Authors give the name of the algorithm as OApriori [7].

Algorithm [4] is based on transaction reduction in which the attribute SOT (size of transaction) is used.As opposed to apriori methodology authors [18] has enhanced joining plus trimmed process. In attempt to remedy the low productivity and effectiveness of the algorithm induced by its producing loads and loads of candidate sets as well as searching the transaction database recurrently, it conducted a study the trimming optimization and operationlessening measures, and on this grounds, the overviews proceduregrounded on trimming improvement and transaction lessening is placed. As per performance analysis in the recreated environment, by applying the effective procedure, the amount of recurrent item sets is considerably fewer and the operation period is drastically condensed along with, the efficiency is boosted then eventually the method is upgraded.

In [14] Authors offer the nomenclature of the technique named APRIORI-IMPROVE founded on the shortcomings of Apriori. APRIORI-IMPROVE method gave improvements on 2-items creation, operation simplification so on and so forth. APRIORI-IMPROVE employs hashing format to construct L2, using an excellent longitudinal data format and enhanced approach of storing save the spacetime. The benchmark analysis demonstrates that APRIORI-IMPROVE is way more efficient over Apriori.

In [19] authors introduce an enhanced algorithm relies on the quintessential Apriori algorithm, wherein Trimmed methodology- is tailored to optimize performance of trimming in attempt to lessen the formation of frequent items, and start taking Transaction Reducing way to mitigate the dimensions of the database table to be processed.

IV. COMPARISON

Different methods, advantages and limitations of improved Apriori algorithm are summarized in the table below.

Sr. No	Title	Method	Advantages	Drawbacks
1	Research and Improvement of Apriori Algorithm[5]	Database: Item-tid form. Dynamic array vector and map set are used as storage structure. Join: L_{k-1} with L_1. Support: Use intersection	Repetition of comparison of (k-2) are avoided. Database scanning avoided. Ascending order of tid-list.	Generation of candidate sets are increased.
2	Enhanced BitApriori Algorithm[20]	BitApriori Equal Support Pruning(ESP) Binary String XOR Operation	Database is scanned only twice. ESP diminishes height of the tree.	May suffer scarcity of memory when scanning huge database. Repeated traversing of the tree.

Sr. No	Title	Method	Advantages	Drawbacks
3	Proposed Algorithm for Frequent item set generation[2]	Database: Adjacency matrix. Pruning: Sorted table and adjacency matrix are pruned. Candidate: Contains item-set related to L_{k-1}.	Decreased usage of memory. Number of items to get scanned gets reduced.	Few item-sets are missed when generating.
4	A Method to Optimize Apriori Algorithm for Frequent Items Mining[7]	Database: Item-tid form. Connection steps: Transactions form L_{k-1} that do not satisfies the min sup should be removed. Support: use intersection.	Database is scanned only once. Checking of (k-1) subsets are avoided.	
5	An Efficient Filtration Approach for Mining Association Rules[8]	Filtration approach is applied to generate frequent and infrequent datasets.	Best possible item-sets are generated.	Additional memory is required to store the infrequent datasets.
6	Improved Filtration step for Mining Association Rules[9]	Filtration steps: Only (k-1) infrequent item-sets are checked.	Best possible item-sets are generated. No extra effort is required in filtration process.	Additional memory is required to store the infrequent datasets.
7	Improving Efficiency of Apriori Algorithm Using Transaction Reduction[4]	Traction reduction strategy: Delete transactions which are less than k.	Traction scanning is reduced.	Have to manage database each time after generation of L_k.
8	An efficient algorithm for frequent item-sets in data mining[6]	BitApriori. Equal Support Pruning (ESP). Binary string. Bitwise AND operator.	Database is scanned only twice. No candidate generation. ESP reduces tree height. Support count can be easily calculated. Only frequent item-sets are generated.	May suffer scarcity of memory when scanning huge database because of BitApriori.
9	Mining AssociationRules Based on an Improved Apriori Algorithm[18]	Exclusion of item-sets not comprising candidate itemset during the computation for the very same.	Number of candidate item-sets are reduced. Efficiency gets higher with increasing value of k.	
10	An Improved Apriori Algorithm Based	Eliminate operation relying happening length parameter to minimize transaction volume and	Reduced scanning of database. Non-frequent item-sets are	During generation of k>2 item-sets, additional memory is

Sr. No	Title	Method	Advantages	Drawbacks
	on Pruning Optimization and Transaction Reduction[19]	also utilization of temporary table to trim and choose frequent itemset.	deleted.	required to store the temporary table.
11	An Effective Hash-Based Algorithm for Mining Association Rules[13]	Candidate item-set generation is done by a hash method during the initial iterations. Reduces the transaction database size by pruning.	Considerable amount of reduction of database size. Scans database twice and then scans D_k. Small transaction scans.	Time complexity is high due to the hashing process and generation of item-sets and support count.
12	Sampling Large Database for Association Rule[17]	Examining the representative selection for identifying common item-sets rather than complete dataset employing lower bit support cutoff.	Saves time due to less scanning. Efficiency is the highest.	Partitioning may lead to loss of some item-sets.

V. CONCLUSION

Being one of the most significant and most used algorithms, Apriori still have some drawbacks like generation of more candidates and multiple scanning of the database. In order to overcome these issues, numeral number of procedures and improvements has been recommended. All these various propositions may have been successful to solve one problem or the other but also has left behind some other minor issues.

REFERENCES

[1] Agrawal R and Srikant R., "Fast algorithms for mining association rules", In Proceedings of the international conference on very large data bases (VLDB'94), Santiago, Chile, 1994.

[2] Archana Singh and Kyoti Agarwal, "Proposed algorithm for frequent item set generation", IEEE, 2014.

[3] Feng WANG and Yong-hua LI, "An improved Apriori algorithm based on the matrix", IEEE, International Seminar on Future BioMedical Information Engineering, pp. 152-155, 2008.

[4] Jaishree Singh, Hari Ram and Dr. J.S. Sodhi, "Improving Efficiency of Apriori Algorithm Using Transaction Reduction", International Journal of Scientific and Research Publications, Vol.3, Issue 1, pp.1-4, 2013.

[5] Jiaoling Du, Xiangli Zhang, Hongmei Zhang and Lei Chen, "Research and Improvement of Apriori Algorithm", IEEE sixth International Conference on Science and Technology, pp.117-121, 2016.

[6] Jiemin Zheng, Defu Zhang, Stephen C. H. Leung, and Xiyue Zhou, "An efficient algorithm for frequent itemsets in data mining", IEEE, 2010.

[7] Ke Zhang, Jianhuan Liu, Yi Chai, Jiayi Zhou and Yi Li, "A Method to Optimize Apriori Algorithm for Frequent Items Mining", IEEE Seventh International Symposium on Computational Intelligence and Design, pp.71-75, 2014.

[8] Lalit Mohan Goyal and M. M. Sufyan Beg, "An Efficient Filtration Approach for Mining Association Rules" IEEE, International Conference on Computing for Sustainable Global Development, pp.178-185, 2014.

[9] Lalit Mohan Goyal and M. M. Sufyan Beg, "Improved Filtration Step for Mining Association Rules", IEEE, 2014.

[10] Mannila H., Toivonen H., Verkamo A., "Efficient algorithm for discovering association rules", In: AAAI Workshop on Knowledge Discovery in Databases, pp.181-192, 1994.

[11] Mingzhu Zhang and Chang zheng He, "Survey on Association Rules Mining Algorithms", Springer, Advancing Computing, Communi, Control and Management, LNEE 56, pp.111-118, 2010.

[12] O.Jamsheela and Raju.G, "Frequent Itemset Mining Algorithms: A Literature Survey", IEEE International Advance Computing Conference (IACC), pp.1099-1104, 2015.

[13] Park J.S., M.S. Chen and P.S. Yu, "An effective hash based algorithm for mining association rules", ACM SIGMOD, pp.175-186, 1995.

[14] Rui Chang and Zhiyi Liu "An Improved Apriori Algorithm" IEEE, International Conference on Electronics and Optoelectronics (ICEOE), vol.1, pp.476-478, 2011.

[15] Savasere A., E. Omiecinski, S.B. Navathe, "An efficient algorithm for mining association rules in large databases", in Proceedings of 21th International Conference on Very Large Data Bases (VLDB'95), 1995.

[16] Surbhi K. Solanki and Jalpa T. Patel, "A Survey on Association Rule Mining", Fifth International Conference on Advanced Computing & Communication Technologies, pp.212-216, 2015.

[17] Toivonen H., "Sampling large databases for association rules", In Proceedings of 22nd VLDB Conference, India, pp.1-12, 1996.

[18] Yanfei Zhou, Wanggen Wan, Junwei Liu, Long Cai**,** "Mining Association Rules Based on an Improved Apriori Algorithm", IEEE, pp.414-418, 2010.

[19] Zhuang Chen and Shibang Cai, Qiulin Song and Chonglai Zhu, "An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction" IEEE, pp.1908-1911, 2011.

[20] Zubair Khan, and Faisal Haseen, "Enhanced BitApriori Algorithm: An Intelligent Approach for Mining Frequent Itemset", SPRINGER, Vol.1, pp.343-350, 2015.

[21] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques" 2nd Edn; Morgan Kaufmann Publishers, ELSEVEIR 2006.