Some observations of Improved Apriori Algorithms

Pankaj Singh
Research Scholar
Dept. of Computer Science
BHU, Varanasi
E-mail: pankajbhumca07@gmail.com

Dr. Rakhi Garg
Assistant Professor
Dept. of Computer Science
MMV, BHU, Varanasi
garg.rakhinetq@gmail.com

Dr. P. K. Mishra
Professor
Dept. of Computer Science
BHU, Varanasi
pkmisra@gmail.com

Abstract: Frequent item set generation play a vital role in Data Mining research. Apriori Algorithm is a basic algorithm of Association Rule Mining but due to large number of database scans and candidate generation, its performance is low with respect to time complexity. To overcome these shortcomings, a lot of works have been done in the designing of improved Apriori algorithm.

In this paper we have reviewed the improved Apriori algorithm that can help scientists, researchers working in this area to identify the problem so that enhancement in the performance of algorithm can be done. Moreover, we have also discussed the merits and demerits of different Apriori algorithms.

Keywords- Association Rule, Frequent Pattern Mining, Apriori Algorithm, Improved Apriori.

I. INTRODUCTION

Data Mining is a way of obtaining hidden patterns from huge amount of data in a database [6]. There are many applications of data mining like customer retention, market analysis, production control and fraud detection. Data mining methods can be categorized into classification, clustering, association rule mining, sequential pattern discovery, regression etc. Among these methods, association rule mining is very important which results in generating strong association rules. Association rule mining was first planned by R. Agrawal [1] which aims at finding frequent itemsets from a set of transactional databases. The various association rule mining algorithms were developed e.g. Apriori, FP-Growth, Direct hashing and pruning (DHP), Apriori T.id etc. The generation of frequent itemsets and association rules depends on the support and confidence value in all association rule mining algorithms.

Pankaj Singh is working as an Assistant Professor in Computer Education in Faculty of Education, Banaras Hindu University. E-mail: pankajbhumca07@gmail.com.

The definitions of support and confidence in term of probability are:

Support (A $\rightarrow$ B) =P (AUB), Confidence (A $\rightarrow$ B) =P (B/A), where A & B are set of items.

The rule that meet the condition of minimum support (min_sup) and minimum confidence (min_conf) values are known as strong association rules. The itemsets which appears in the data set frequently are known as frequent itemsets. If the support value of itemsets A is more than or identical to minimum support threshold value, then itemsets A is called frequent itemsets. [6].

The generation of frequent itemsets and association rules depends on the support and confidence value in all association rule mining. The efficiency of algorithm depends upon the times it takes to generate frequent itemsets and association rule which further depends upon the threshold values of support and confidence.

This paper is divided into four sections including Introduction. Section two defines related work on Apriori Algorithm, Section three represents the improved Apriori Algorithm while Section four presents overall conclusion.

1. RELATED WORK

Agrawal et al proposed AIS algorithm [1] for generating frequent itemsets. In the AIS algorithm, frequent itemsets are generated through iterations on scanning the database. The iteration terminates when no new frequent itemset is derived. After reading a transaction in the k^{th} iteration, the AIS algorithm computes the candidate k itemsets by first deriving a set of (k-1) itemsets which contains itemsets that are both in the

frequent (k-1) -itemsets and in the transaction. Ayres. J [2] introduced an effective pruning mechanism called Depth First Search to extract the sequential pattern in huge database. This method defines the database in vertical bitmap format (VBF) with efficient support counting. For all item in the dataset, a vertical bitmap (VB) is formed by which each data set transaction is represented as a bit. The value for items is set based on the item present in the transaction. The efficient support counting and candidate generation is obtained by partitioning the bitmap. Changsheng Zhang and Jing Ruan [3] have worked on the improvement of Apriori algorithm by applying dataset reduction method and the reduction of I/O spending. They have applied the modified algorithm for instituting cross selling policies of the trade industry and to improve the sales performance. Chen. J [4] presented a BISC (Binary item set Support Counting), which is responsible for efficient mining of frequent item set. According to this algorithm, support of all item set in a database is derived with respect to direct supports. The context BISC transforms the database transaction into binary format. The memory consumption and the cost of support updating are minimized by integrating the two related techniques namely, BISC1 for one stage and BISC2 for two stages. This technique is both time and space efficient because of BISC in conjunction with projection techniques that reduces the branching factor of database projection and a maximum depth.

Dongme Sun, Sheohue Teng [5] have presented a new technique based on forward and reverse scan of database. It produces the frequent itemsets more efficiently if applied with certain satisfying conditions. Hanbing Liu [7] proposed a novel association rule algorithm (ARM) called Association Rule Mining Based on Boolean Matrix algorithm which converts a transaction database into a Boolean matrix. It scans the transaction database once, does not produce candidate itemsets, and takes on the Boolean vector "relational calculus" to discover frequent itemsets. In addition, it stores all transaction data in bits, so it needs less memory space and can be applied to mining large database. J. S. Singh et al [8] developed a TR algorithm which reduced the searching time by deleting unnecessary transaction records and the redundant generation of sub items during pruning of the candidate item sets, which can form directly the set of frequent itemsets and eliminate candidates having a subset that is not frequent (L) but it has overhead to manage the new database after every generation of L_k. So, an approach is required that should minimize the number of database scans. Kavitha. K [9] presented an efficient transaction reduction technique named TR-BC to extract the frequent pattern based on bitmap and class support as a substitute of only item support. Moreover, the database storage is condensed by using bitmap that notably compress the number of database scan. The rules are condensed by horizontal and vertical transaction (HTVT) and then finally joint rules are generated by deleting the redundancy. Ramaraj. E [10] proposed an algorithm called TR Apriori that manages its performance even at relative low supports. The benefit of TR Apriori includes interactive mining with different supports, less execution time and storage of infrequently used item to improve the size of the query data. Sixue Bai and Dai [11] have presented a method called P-matrix algorithm to generate the frequent itemsets. It is found that the P-Matrix algorithm is more efficient and fast algorithm than Apriori algorithm to generate frequent itemsets. Wanjun Yu, Xiaochun Wang [12] have presented an algorithm called as Reduced Apriori Algorithm with Tag , which increases the performance of Apriori algorithm by decreasing the number of frequent itemset produced in pruning operation, by pertaining transaction tag method. Zhi Lin, Guoming Sang, Mingyu Lu [13] proposed a vector operation based method. The algorithm finds the association rule more efficiently and requires only one database scan to find all the frequent itemsets. The merits and demerits of all algorithms are discussed in Table-1 shown below.

3. IMPROVED APRIORI ALGORITHM

A lot of work has been done to improve the performance of the traditional Apriori algorithm and some of the factors are shown in figure-1.

3.1 Reducing redundant pruning operation and increasing efficiency of support calculation

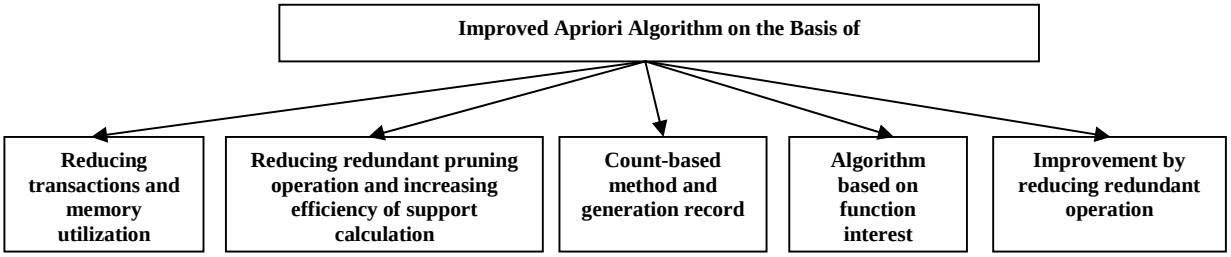
Wanjun Yu, Xiaochun Wang, Fangyi Wang, Erkang Wang and Bowen Chen [2] has proposed an algorithm known as Reduced Apriori Algorithm with Tag (RAAT). RAAT uses Apriori-gen operation to form candidate 2-itemsets which results in diminishing the pruning operation. RAAT

also follows the concept of tag to increase the speed of support calculation. As a result, RAAT shortens the time and improves efficiency. The experimental result shows that RAAT performs

Table -1: Summary of Apriori Algorithm

Authors Name	Algorithm's Name	Description	Merits/Demerits
Agrawal et al [1]	AIS	Frequent Itemsets are generated through iterations on scanning the database	Less data support, very less accuracy and slow speed in initial phase.
Ayres. J [2]	Depth First Strategy	The database is in vertical bitmap layout with efficient support counting	Apply a depth first traversal of the search space.
Zhang. Changsheng and Ruan . Jing[3]	IAA	Applying dataset reduction method and by reducing the I/O spending	Reduce the redundant operation while generating frequent itemset.
Chen. J [4]	BISC (Binary Item set Support Counting)	Support of all item set in a database is derived with respect to direct supports. The context BISC transforms the database transaction into binary format.	Reduce the total number of database projection , efficient at deriving itemset support based on direct support and memory consumption is O(km) where k-maximum projected depth, m-average size of projected database
Dongme Sun, Sheohue Teng [5]	IAA	Presented a new technique based on forward and reverse scan of database	It avoids producing candidate itemset. Also saves time and space.
Hanbing Liu [7]	Association Rule Mining supported on Boolean Matrix algorithm	Converts a transaction database into a Boolean matrix. It searched the transaction database once and does not generate candidate itemsets. It accepts the Boolean vector "relational calculus" to apprehend frequent itemsets	Scan the database only once does not produce candidate itemset. It stores all transaction data in bits.
Jaisree Singh et al [8]	Transaction Reduction Algorithm	Reduced the scanning time by deleting redundant transaction records as well as decrease the redundant generation of sub-items during pruning of the candidate itemsets, which can form openly the set of frequent itemsets and remove candidate having a subset that is not frequent	This algorithm not only optimizes the algorithm by reducing the size of the candidate set of k-itemset , C_k but also reduces the I/O spending by cutting down transaction records
Kavitha. K[9]	TR-BC	To extract the frequent pattern created on bitmap and class labels. The process decreases the rule creation by counting the item support and class support as a substitute of only item support	Fewer number of database scanning is done that reduces number of rules, time and space complexity.
Ramaraj.E [10]	TR Apriori	Manages its performance even at relative low supports	Interactive mining with different supports, fasten the execution time
Sixue Bai and Xinxi Dai [11]	P-matrix algorithm	To generate the frequent itemsets faster than Apriori	Scan for database once to obtain the binary pattern matrix and transform its ranks
Wanjun Yu, Xiaochun Wang [12]	Reduced Apriori Algorithm with Tag (RAAT)	Reducing the number of frequent itemset created in pruning operation, by concerning transaction tag method.	RAAT algorithm works faster than traditional for same support and with the support increased , the operating efficiencies of two algorithm has been enhanced
Zhi Lin, Guoming Sang, Mingyu Lu [13]	Vector operation based method	Requires only one database scan to find all the frequent itemsets.	Transactional Database need to be scanned only one time to generate the Boolean matrix which is stored in bit mode, so the memory space is greatly scanned

Figure-1: Criterion for improving Apriori Algorithm



3.2. Reducing transactions and memory utilization

Jaishree Singh, Hari Ram, Dr.J.S.Sodhi [8] has presented an improved Apriori Algorithm (IAA) to conquer the limitations of traditional Apriori Algorithm. The traditional Apriori algorithm scans the database many times. If database contains large number of records, it takes large amount of time to scan the database which results in increasing I/O cost. The improved Apriori Algorithm reduces the scanning time by eliminating the transactions containing irrelevant records. It uses the concept of attribute named as Size of Transaction (SOT) which contains the number of items exists in specific transaction. By comparing improved Apriori Algorithm with traditional Apriori algorithm, it was found that improved Apriori Algorithm is better on the basis of efficiency and optimization [8].

3.3 Count-based method and generation record

Huan Wu, Zhigang Lu, Lin Pan, Rongsheng Xu and Wenbao Jiang [6] has proposed an improved Apriori algorithm (IAA) for association rule mining. The IAA overcomes the restrictions of original Apriori algorithm. The IAA introduces a new count based approach which is used to extract the redundant candidate itemsets and utilizes creation record to shrink the scanning time of database. The IAA meets the various challenges correlates with association rule mining such as reducing I/O cost, improving efficiency and increasing processing speed. From the experimental results, it was proved that IAA is better than original Apriori algorithm because IAA counts each candidate itemsets once. But **C-R problem** exists in IAA which could not be solved where C represents condition item sets and R represents result item sets.

3.4 Algorithms based on function interest

The classical Apriori algorithm generally focuses on only two aspects: minimum support and minimum confidence to generate strong association rule. There may be a chance that sometimes it is necessary to determine strong association rules for making decisions and sometimes less strong rules are required. To accomplish these conditions, *WEI-MIN MA and ZHU-PING LIU* [8], proposed two revised algorithms based on Apriori: AMS (Algorithm for mining stronger association rules) and AMLS (Algorithm for mining less strong association rules) which focus on three aspects: minimum support, minimum confidence and minimum interest. These algorithms works in the form of matrix to decrease the scanning time of database. On the basis of comparison of traditional Apriori algorithm with AMS and AMLS, it was proved that AMS and AMLS are better than traditional Apriori algorithm.

3.5 Improvement by reducing redundant operation

Yanfei Zhou, Wanggen Wan, Junwei Liu and Long Cai [9], described an improved Apriori algorithm. This improved Apriori algorithm consists of three segments: First is decreasing number of judgments during the time of generating frequent candidate itemsets and secondly, pruning frequent itemsets. Finally, optimize the database. The improved Apriori algorithm was compared with classical Apriori algorithm on the basis of different support degree, different number of trading services and different number of items.

From this comparison, it was proved that improved Apriori algorithm improves performance, increases efficiency, and reduces the redundant operation while producing frequent itemsets and strong association rules.

4. CONCLUSION

After studying many algorithms based on Apriori improvement, we conclude that the entire algorithm reduces the number of database scan and generates less number of candidate set and effective association rule which is based on support and confidence value. These algorithms are better than traditional Apriori algorithm but many improvements can be done in generating frequent itemsets. In this paper, we have seen that the improved algorithm not only optimizes the algorithm of dropping the size of the candidate set but also reduces the I/O costs by cutting down transaction records in the database. Also it has overhead to maintain the new database after every creation. Apriori increases from view of time consumed and whenever the value of minimum support increases, the breach between improved Apriori and the classical Apriori decreases from view of time consumed. Some algorithms scan the database only once and reducing the system I/O cost by using array to store data. This area has lot

of scope due to rapid growth of data day by day. The merits and demerits of each improved Apriori algorithm discussed in this paper can help researchers and scientist to design an efficient algorithm.

REFERENCES

1. Agrawal. R. And Srikant. R., "Fast Algorithms for mining Association Rules", Proceedings of 20th International Conference of very large databases, 1994, pp. 487-499.
2. Ayres.J, Flannick.J, Gehrke.J, and Yiu.T, "Sequential pattern mining using a bitmap representation," inFlannick.J, Gehrke.J, and Yiu.T, "Sequential pattern mining using a bitmap representation," in proceedings of the eight ACM SIGKDD international conference on knowledge discovery and data mining 2002, pp. 429-435.
3. Changsheng Zhang and jing Ruan "A Modified Apriori Algorithm with its application in Instituting Cross-Selling strategies of the Retail Industry," in Proc of International Conference on Electronic Commerce and Business Intelligence, 2009, pp. 515-518.
4. Chen.J and Xiao.K, "BISC: A bitmap itemset support counting approach for efficient frequent itemset mining," ACM Transaction on Knowledge Discovery from Data (TKDD), 2010. VOL. 4, P. 12.
5. Dongme Sun, Sheohue Teng, " An algorithm to improve the effectiveness of Apriori Algorithm," in proc. Of 6th ICE Int. Conf. On Congnitive informatics, 2007, pp. 385-390.
6. Han.J, Kamber.M, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Book, 2000.
7. Hanbing Liu and Baisheng Wang "An Association Rule Mining Algorithm Based on a Boolean Matrix", Data Science Journal,2010 Vol (6), supplement 9.
8. Jaisree singh, Hari Ram, Dr.J.S.Sodhi "Improving efficiency of Apriori Algorithm using Transaction Reduction" IJSRP, Vol(3), 2013. ISSN 2250-3153.
9. Kavitha.K, Dr.E.Ramaraj,"Efficient Transaction Reduction in Actionable Pattern Mining for High Voluminous Datasets based on Bitmap and class Labels", IJCSE, Vol.5 No.07 jul 2013. ISSN: 0975-3397.
10. Ramaraj. E, K. RameshKumar, N.Venkatesan, "A Better Performed Transaction Reduction Algorithm for Mining Frequent Itemsets from Large Voluminous Database", Proceeding of the 2nd National Conference, Computing for Nation Development, February 08-09, 2008.
11. Sixue Bai, Xinxi Dai, "An efficiency Apriori algorithm: P_matrix algorithm," First International Symposium on Data, Privacy and E Commerce, 2007, pp.101-103.
12. Wanjun Yu, Xiaochun Wang., "The Research of Improved Apriori Algorithm for Mining Association Rules," in Proc. Of 11th IEEE International Conference on Communication Technology Proceedings, 2004 pp. 513-516.
13. Zhi Lin, Guoming Sang Mingyu Lu "A Vector Operation Based Fast Association Rules Mining Algorithm," in Proc. of Int. Joint Conf. on Bioinformatics, System Biology and Intelligent Computing, 2009, pp. 561-564.

Pankaj Singh is a Ph.D. Scholar at Department of Computer Science, Banaras Hindu University and working as an Assistant Professor in Computer Education for two years in Faculty of Education, Banaras Hindu University Banaras Hindu University, India. He received his M.C.A. from the same Department in 2010. His research interest includes Data Mining Algorithms, Parallel and Distributed Computing and Big Data. He has authored two research articles published in journal and conference proceedings.

Rakhi Garg received her M.Sc. in 1997 and Ph. D. in 2012 from Department of Computer Science, Banaras Hindu University, India. She has more than 15 years of teaching experience at various Institutions. Currently she is a Senior Assistant Professor and In-charge at Department of Computer Science, Women's College, Banaras Hindu University since 2007. Her research interests include Data Mining, Web Mining, Cluster Computing and High Performance Computing. She has authored more than 10 research articles in journals and conference proceedings. She has organized and attended a number of research workshop and training programs and delivered invited lectures in various conferences.

P. K. Mishra is a Professor and Head, Department of Computer Science, Banaras Hindu University, India. He is also a Principal Investigator of the research projects at DST Centre for Interdisciplinary Mathematical Sciences, Banaras Hindu University. His research interests include Parallel and Distributed Computation, Computational Complexity, Parallel and Clustered Data Mining, High Performance Computing and VLSI Algorithms. Prof. Mishra has more than 70 publications in conference proceedings and journals. He is a reviewer and editor of various journals and senior member of IEEE. He has organized a number of conferences and research workshops and delivered invited talks in a number of International conferences and workshops.