International Journal of Computer Science & Communication
Volume 6, Issue 2, April-September 2015
ISSN: 0973-7391 Impact Factor: 2.5 Available at www.csjournals.com

Comparative Analysis of Apriori Algorithm based on Association Rule

Apurwa Sahu, Mradul Dhakar, Pushpi Rani
Department of CSE, ITM University Gwalior. INDIA
apurwa.670@gmail.com, mradul.dhakar.cse@itmuniversity.ac.in, pushpirani@itmuniversity.ac.in

Abstract: Apriori algorithm is the first or the traditional algorithm of association rules which find out all the frequent item-sets from transaction database. This paper presents a review report about the various Apriori algorithm proposed by different researchers. The existing algorithms have some limitations like, time complexity, space complexity and no. of passes. Therefore, it is needed to design an alternative approach for association rules mining to enhance the Apriori algorithm and reduce its time complexity.
Keywords: Data mining, frequent pattern, support, Confidence, Association Rule, Apriori Algorithm, soft set.

1.INTRODUCTION:

Data mining is used for "extracting knowledge from large amounts of data". Association rule mining (ARM) is best one data mining technologies. ARM is a process for finding relations between data items in datasets. Association rule mining has been proven to be a successful technique for extracting knowledge. Frequent patterns are patterns item-sets that appear frequently in data set. For Example, bread and butter is a set of items which is appeared frequently together in a transaction [1]. Main motive to search frequent pattern in any data set is for analysis of any data for example if we want to find out which product is frequently purchased by customers in supermarket transaction database then we can easily check which product is frequently exists in daily transaction data of supermarket [4].The next section contains the brief introduction about association rule mining followed by Apriori algorithm. Further there is a description of existing research on Apriori algorithm. At the end of section we have compared them on the basis of time complexity.

1.1 ASSOCITAION RULE MINING:

Association rule mining (ARM) is used for finding the frequent patterns between the item-sets. Its goal is extracting interesting association rules, frequent patterns among item-sets in the database. For Example, In a Laptop store in Delhi, 90 % customer who purchased a laptop, they also buy a mouse. The statement of ARM problem was firstly specified by R. Agrawal. Let E = E_1, E_2... E_n be a set or group of n different attributes, B be the transaction such that B ⊆ E, G be a database with different transactions B_s. An association rule is defined as M ⇒ N, where M, N ⊂ E are sets of items, and M ∩ N = φ. M is called antecedent. N is named consequent. The rule means M implies N. Support and Confidence are the two basic measures for association rules. These measures are used for the evolution of association rule interestingness. There are two thresholds that is minimum support and minimum confidence which can define by the users.

Support: The support is defined as the probability of task with relevant data transactions for which the pattern is true [2].

$$\text{Support } (A \cup B) = P(A \cup B)$$

Confidence: The confidence is defined in terms of measure of certainty associated with each discovered pattern [2].

$$\text{Confidence } (A \cup B) = P(B|A)$$

If rules that satisfy both a minimal support threshold (and minsup) and a minimal confidence threshold (or minconf) are called strong association rules [1].

1.2 APRIORI ALGORITHM:

Apriori algorithm employs the level wise approach or width search method, it is included all the frequent itemsets [3]. Association rules is guided by two parameters: support and confidence. An association rule is returned by Apriori if user defined threshold values satisfies its minimum support and confidence values. The output is grouped by minimum confidence. Rules are ordered by support only if they have the same confidence. So Apriori favors more confident rules than any other algorithm and characterizes them as more interesting rules. The Apriori Mining process has two major steps. The first one is generating frequent item-sets of the Apriori algorithm. If item-sets with at least minimum support were considered then this step can be seen as support based pruning. The second step is the generation

of rules, in which confidence based pruning is applied. Rule discovery use straightforward mechanism [2, 4].

2. OVERVIEW OF RECENT RESEARCH IN APRIORI ALGORITHM:

D.N.Goswami et.al [4] suggested three approaches, Record filter, Intersection and proposed algorithm, based on traditional Apriori Algorithm. These approaches enhanced the performance of time and no of data passes. The Record Filter approach counts only the support of candidate set in the transaction record. It consumes very less time in compare with traditional apriori algorithm. Intersection Algorithm enhanced the efficiency, memory management and reduces the computation cost of Apriori Algorithm. This approach was more designed for vertical data format, so it removed the limitations of horizontal data format used in Apriori. The Proposed algorithm merged the concept of both algorithm that is Record filter algorithm and Intersection algorithm which is based on set theory concept of intersection with the record filter approach. This paper concluded that the Record Filter approach is better than Classical Apriori Algorithm, Intersection Approach is better than Record Filter approach and at last, Proposed Algorithm is much better than other frequent pattern mining algorithm. Proposed Apriori Algorithm takes less time than the Classical Apriori Algorithm.In this algorithm, researchers have the key ideas for reducing time complexity. Proposed approach is really worked in saving the time complexity of large database.

Yihua Zhong [5] suggested that association rule (AR) is an important model in data mining. However, traditional association rules are based on the support and confidence metrics. Weighted Rule Algorithm had proposed by the author and it can be reduced the large number of unnecessary association rules and mined interesting negative association rules.

Rachana Somkunwar [2] had proposed an enhanced version of Apriori Algorithm. It is focused on four characteristics: first, it prepared a data and then chooses the desired data, second, it produced item-sets that is used to decide the rule constraints for knowledge, third one is mined k-frequent item-sets using the new database and, fourth is produced the association rule that sets up the knowledge base and offer better results. In this paper, author had used the mapping-table to convert the items into compressed data set. To count the support, HASH_TREE had been used, which disperses the matching of candidate item-sets to reduce the time complexity of algorithm. This approach used HASH MAPPING TABLE and HASH_TREE, to optimized both space and time complexity.

Jiao Yabing [3] introduced that Apriroi Algorithm is the Classic Algorithm of association rules in data mining. The author proposed an Optimized Algorithm which improved the efficiency of the Classical Apriori Algorithm. In [3], candidate key was compared with support level once it was found, and generate L_{k-1} that has been pruned from item set less than the support level. Further, this paper is described the optimized algorithm, decreases the number of connecting items sets, and the number of candidate items also declined. It improves the efficiency of ARM. As, it reduces the number of candidate items set and also save time complexity.

Komal Khurana et.al [7] presented a comparative analysis among different association mining algorithms. Five association rule mining algorithms are presented which are compared with each other, that is, AIS, SETM, Apriroi, AprioriTid and AprioriHybrid. These algorithms had compared on the basis of several factors like type of data set, support count, rule generation, candidate generation and similar factors. AIS was first algorithm. It consists two phases-first phase described the generation of the frequent item-sets and in second phase constituted the generation of conflict and frequent association rule. The drawback of this approach is multiple passes over the database. Second algorithm was SETM; it was motivated by the desire to use SQL to compute large item-sets. It is same as AIS but count the item set at the end of the pass. Third algorithm was Apriori Algorithm which generated lesser candidate set of item-sets for testing in every database pass. Fourth algorithm was AprioriTid Algorithm this used the apriori-gen function. This function determines the candidate itemsets before the pass begins. The fifth algorithm was AprioriHybrid Algorithm. This approach used different algorithm for different passes. AprioriHybrid is better than all other algorithms, it reduce the speed and improve the accuracy.

Sujatha Kamepalli et.al [7] introduced a method based on Apriori Algorithm to find out infrequent item-sets and non-present item-sets from transactional database with one database scan. In Apriori Algorithm the frequent patterns are found and infrequent patterns are removed, a pruning technique is based on Apriori principle that is if an item set is found to be infrequent then all its super sets are also infrequent. Further, this approach find out infrequent and non-present item sets based on Apriori principle but in this super sets are not to be pruned like Apriori, and they are considered into the solution as in frequent k-item sets.

Yuanyuan Zhao [8] suggested that the Negative association rules became a focus in the data mining field. Negative association rules (NARs) were useful in market-basket analysis to identify products that complement each other. The negative association rules mostly consisted in the infrequent items. It was proved that the number of the negative association rules (NARs) from the infrequent items is larger than those from the frequent. Table 1 shows a review of work done by different authors. Further, we have analyzed the experimental result described in [1],[4],[9],[10]. The comparative analysis of these results has been shown in Table 2 and figure 1 depicts its graphical representation.

3. LIMITATIONS OF APRIORI:

- Apriori algorithm is not good for large database.
- This algorithm only defines the presence and absence of an item.
- This algorithm is allowed uniform minimum support threshold.
- It is limited for small database.
- More scanning is needed of transaction database for calculating frequent item.
- Generation of candidate item-sets and support counting are expensive.

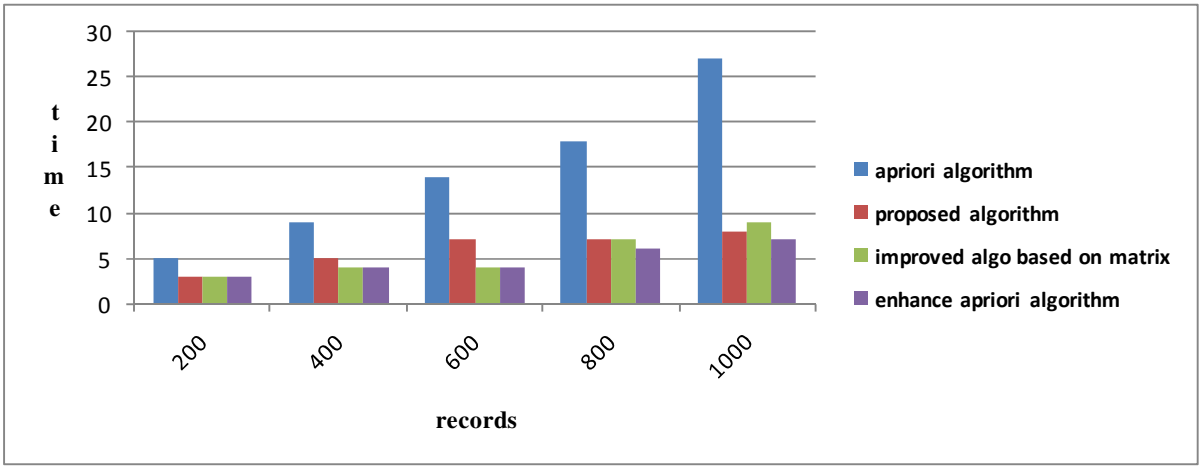
Table 1: Summary of literature work on Apriori algorithm in chronological order

S.NO.	AUTHORS	YEAR	METHODS
1	J.Han and Kamber[1]	2007	Traditional Apriori Algorithm
2	Yuanyuan Zhao et.al[8]	2009	Weighted Negative Association Rules Based on Correlation from Infrequent Items
3	D.N.Goswami et.al[4]	2010	Record filter, Intersection and Proposed Algorithm
4	Yihua Zhong et.al[5]	2012	Weighted Rule Algorithm
5	Rachna Somkunwar[2]	2012	HASH MAPPING TABLE and HASH_TREE
6	Jiao Yabing[3]	2013	Optimized Algorithm of Apriori Algorithm
7	Komal Khurana et.al[6]	2013	Comparisonc between AIS,SETM,Apriroi,AprioriTid and AprioriHybrid
8	Sujatha kamepalli et.al[7]	2014	Mining Infrequent and Non-Present Item Sets based on Apriori Algorithm

Table 2: Experimental results of different algorithms

S.no.	1	2	3	4
Algorithms / Records	Apriori algorithm[1] (in sec.)	Proposed algorithm[4] (in sec.)	Improved algorithm based on matrix[9] (in sec.)	Enhancement in Apriori Algorithm[10] (in sec.)
200	5	3	3	3
400	9	5	4	4
600	14	7	4	4
800	18	7	7	6
1000	27	8	9	7

Figure1: Graphical representation of Table 2.



4. PROPOSED METHODOLOGY:

The comparative result analysis of existing algorithm has been shown in figure 1. The aim of research is to improve the efficiency of existing algorithms by reducing its time complexity. For this we would use the soft set-theory approach with the existing one. Soft set is a way for mining of data by association rule from transactional dataset. This approach transforms all the transactional data sets as a Boolean value in existing information system. At present, work on soft set theory is progressing rapidly and many important results have been achieved. The main advantages of this soft set theory are: there is no limit of parameterization tool and they can deal with Boolean-valued-information system.

5. CONCLUSION AND FUTURE WORK:

To conclude, our survey details several existing approaches which focus on improving the Aprori algorithm. The data are going to complex day by day and in near future in order to handle an explosion on quantity of data we need several approaches to handle database as well as to enhance the Apriori algorithm. Several approaches were discussed in this paper but again lot of research is required to upgrade existing algorithms. In future work, we will present a methodology based on soft set theory by which we will improve the time complexity as well as space complexity.

6. REFERENCES:

[1] J.Han and Kamber," DataMining: Concepts and Techniques", Beijing: China Machine Press, 2007.

[2] Rachna Somkunwar,"A Study on Various Data Mining Approaches of Association Rules", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), vol.2, issue 9, sep.2012.

[3] Jiao Yabing,"Research of an Improved Apriori Algorithm in Data Mining Association Rules", International Journal of Computer and Communication Engineering, vol.2, no.1, jan.2013.

[4] D.N.Goswami et.al,"An Algorithm for Frequent Pattern Mining Based on Apriory", IJCSE, vol.2, no.04, 2010.

[5] Yihua Zhong et.al, "Research of Mining Effective and Weighted Association Rules Based on Dual Confidence," Computational and Information Sciences (ICCIS), Fourth International Conference on, vol., no., pp.1228, 1231, 17-19 Aug. 2012.

[6] Komal Khurana et.al,"A Comparative Analysis of Association Rules Mining Algorithms", International Journal of Scientific and Research Publications, vol.3, issue 5, may 2013.

[7] Sujatha kamepalli et.al,"Apriori Based: Mining Infrequent and Non-Present Itemsets from Transactional Data Bases", IJECS-IJENS, vol.14, no.03, June 2014.

[8] Yuanyuan Zhao et.al, "Mining Weighted Negative Association Rules Based on Correlation from Infrequent Items," Advanced Computer Control, ICACC '09. International Conference on, vol., no., pp.270, 273, 22-24 Jan. 2009.

[9] Miss. Nutan Dhange et.al ,"Scalable and Efficient Improved Apriori Algorithm", International Journal of Innovative Research in Computer and Communication Engineering,vol.1,Issue 2,April 2013.

[10] K.Geetha et.al,"An Efficient Data Mining Technique for Generating Frequent Item sets",International Journal of Advanced Research in Computer Science and Software Engineering,vol.3,Issue 4,April 2013.