Original Article

A Comparative Analysis of Association Rule Mining Algorithms in Data Mining: A Study

P. Prithiviraj* and R. Porkodi

Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India

Address for Correspondence

Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India.
E-mail: prithiv.mla@gmail.com

ABSTRACT

Data mining is a crucial facet for making association rules among the biggest range of itemsets. Association rule mining [ARM] is one among the techniques in data processing that has 2 sub processes. First, the strategy referred to as finding frequent itemsets and the second method is association rule mining. During this sub method, the principles with the utilization of frequent itemsets are extracted. Researchers developed plenty of algorithms for locating frequent itemsets and association rules. This paper presents the extensive study of various Association Rule mining algorithms and its comparisons. This paper also compared the ARM algorithms based on the merits, demerits, data support and speed.

Keywords: Data mining, KDD, Association Rule Mining, Apriori, AprioriTid, AIS, SETM, Apriori hybrid, FP-Growth.

INTRODUCTION

Mostly, data mining is the expansion of analyzing data as of different perspectives and summarizing it into valuable information-information container exist used to growth expenses, cuts prices, etc. Data mining software is one of a number of systematic tools designed for analyzing data. It consents clients to analyze data since several different dimensions and encapsulate the associations identified. In principle, data mining is the procedure of discovery correlations among loads of fields in huge interactive databases.

At this instant the data mining Techniques established newly includes some main kinds of data mining approaches such as classification, generalization, characterization, clustering, association, evolution, design matching, data imagining and meta-rule directed mining, etc. The techniques for mining, knowledge starting altered types of databases, plus relational, transactional, item focused on, 3-D and dynamic databases by the worthy universal information systems. Possible data mining presentations and some research problems are conferred.

Data mining is famed as some of the main developments of Knowledge Discovery in Database (KDD). Several developers' takings data mining such as a replacement for additional common word, Knowledge Discovery in Database (KDD). Otherwise, further publics give Data Mining As per the main development of KDD. The

KDD processes are shown in Figure 1[10]. Commonly at this time are three developments. One is supposed preprocessing, which is performed previously data mining techniques are functional to the correct data. The preprocessing contains data dusting, assimilation, collection and alteration. The central procedure of KDD is the data mining method, open this process different algorithms are pragmatic to products hidden knowledge. Succeeding that method another procedure called post processing, which estimates the mining produce, allowing to operational requirements and domain knowledge. Concerning the estimation products, the knowledge can be obtainable if the product is acceptable, else we take to write about or all of persons, processes over again until we come to be the acceptable product. The categorically procedures work as trails.

First, we need to clean and integrate the databases. Then the data source could come after altered databases, which may have some inconsistencies and duplications, we must clean the data source by removing those noises or make some compromises. Suppose we have two different databases, changed words are used to mention the similar entity in their schema. While we try to assimilate the two causes we container only select one of them, if we identify that they signify the equal item. And also real world data tend to be incomplete and noisy due to the manual input mistakes. The combined data sources can be deposited in a database.

As not all the data in the database are associated with our mining task, the instant process is to select task associated data from the included capitals and convert them into a format that is ready to be mined. Suppose they want to find which items are often purchased together in a superstore, even though the database that registers the buying history could contain Buyer ID, things subscribed, transaction period, amounts, and number of each products and so on, however for this explicit task our individual need items subscribed. After variety of related data, the database that we are successful to spread over our data mining techniques will be much smaller, consequently the whole process will be more efficient.

Data mining concept are here are two classes of data mining descriptive and predictive. Descriptive mining is to review or characterize common things of data in a data repository. In this descriptive are several techniques clustering, summarization, association rule mining, sequence pattern. Predictive mining is to implement analysis on present data, to create predictions produced arranged the antique data. In this predictive are several techniques they are classification, regression, time series analysis, prediction.

Association rule mining

Association rule mining [ARM] is the one of the best signed and glowing researched methods of data mining, existed initially presented in[3]. Association rule mining is a great resolution designed for substitute rule mining, since its objects to realize entirely rules in data and as a result is able to arrange for a whole depiction of associations in a huge dataset. Present area, yet, two most important problems by way of regard towards the association rule generation. At first problem branches formation the rule quantity and excellence problems. Unknown least provision is set as well as high, the rules concerning intermittent substances that can be of interest to resolution makers will not be initiated. Situation least provision low, however, container cause combinatorial explosion. In other words, else several rules are produced regardless of their interestingness[33]. Several algorithms container be used to realize association rules from data to abstract useful arrays. Apriori

algorithm is one of the greatest extensively used and illustrious techniques for discovering association rules[2,3].

Given a set of relations somewhere each of relatives stays a set of items (itemset), an association rule indicates the form X∪Y, where X and Y stay itemsets; X and Y are called the body and the head, individually. A rule container be calculated by two processes, entitled confidence and support. A ration, support used for association rule X∪Y is the fraction of relations that have both itemset X and Y between all relations. The assurance for the rule X∪Y is the measurement of connections that enclose an itemset Y in the intermediate of the transactions that contain an itemset X. The sustenance signifies the utility of the exposed rules and the assurance signifies the inevitability of the rules.

This paper gives an extensive survey on different association rule mining algorithms particularly Apriori, AprioriTID, Apriori Hybrid, AIS, SETM, FP-growth. The algorithms are analyzed with respect to merits, demerits and its suitability on itemsets. This paper also gives the comparison of algorithms based on speed, accuracy and data support. The paper is organized as follow: Section 1 gives the detailed introduction on data mining and association rule mining. Section 2 discusses the review of literature, section 3 focus on Association Rule Mining Algorithms and its performance. Section 4 discusses on comparison tables and finally the work is concluded in section 5.

Review of Literature

Data mining techniques have designed a partition of useful artificial intelligence, as the 1960s. In the principal periods, main originations in computer systems take directed to the overview of new technologies[9] for Network, established instruction. The explosive development of databases takes to produce an essential to improve technologies that procedure information and information logically. So, Data mining techniques take developed an increasingly main research space[8]. The Apriori algorithm[2]. Employs a bottom-up breadth-first approach to get the huge item set. As it was existing to hold the relational data this algorithm cannot be useful straight to mine compound data. An Apriori based data mining technique is studied at[16]. The initially is mining frequent itemsets with Apriori, and then producing association rules according to the frequent itemsets mined[39]. Apriori is expending circulatory generation for searching frequent itemsets that produces (k+1) itemsets from kitemsets[36]. LI Pingxiang obtainable method explores the database to filter frequent 1-itemsets and then it gets the candidate frequent itemset-2, itemsets-3 up on the way to n-itemset by estimating their possibilities in Equation[23]. The number of database scans required for the task has been reduced from a number equal to the size of the largest itemset in Apriori[2,3].

Algorithms for mining association rules since relational data have existed completed since extended already. Association rule mining was first presented in 1993 by R. Agrawal[3]. Next that several algorithms have been suggested and developed Apriori[19] and FP growth[18]. Additional accepted algorithm is a FP growth algorithm. It expenditures divide-and-conquer method. Initially, it analyses the common items and characterizes the common items in a tree named is frequent-pattern tree. This tree container also exploits as a compressed database. The association rule mining is completed on the compressed database by the use of this FP tree. This signifies that the dataset essentials to be reviewing formerly. Similarly, this algorithm does not need the candidate item set generation. Several modified algorithm and technique has existed suggested by different journalists. Such as FP-tree and COFI based approach is proposed for multilevel association rules. Here except the

FP tree, a new type of tree called COFI- tree is proposed[34].

Attila Gyenesei discussed an important technique of mining association rules for market analysis[4]. The Traditional association rule mining algorithms container only be used to data mining problems with categorical element. Designed for a data mining problem with measureable attribute, it is essential to convert each quantitative attribute into discrete intervals. Agrawal discusses this by way of worthy illustration and implementing association rules method in the purchaser transaction. HUANG Liusheng obtainable an algorithm on BitMatrix, in this algorithm is associated the before identified algorithms Apriori and the AprioriTid algorithms[15].The main task of every association rule mining algorithm is to find out the sets of items that frequently appear together the frequent itemsets.

R. Porkodi presented the rule based approach for constructing gene and protein names dictionary from Medline abstracts that consists of three phases. In the first phase, pre-processing is carried out to remove the inconsistencies from the dataset. In the second phase, the Gene and Protein names are extracted from Medline abstracts using regular expressions and added to dictionary and in third phase the extracted gene and protein names are validated and verified using precision, recall and F-measure. The experimental result shows that the proposed work provides 81% accuracy in identifying Gene and Protein names, which is evaluated and verified using the Precision, Recall and F-Measure. Further it is decided to construct dictionary using any statistical approach rather than rule based approach implemented. The performance of this work may be Compared with the newer one[29].

R. Porkodi presents an efficient framework for extracting relationships between gene ontology terms in biological documents using association rule mining and GO annotations. This may be useful for the biologists to arrive any kind of decisions in their research in gene prediction and identification of deceases in their respective area. This provides set of possible rules for every gene product which may be useful at the time of predicting the gene expression patterns and extracting the relationships between gene products. The associations of various GO terms are grouped by prior biological knowledge which is organized in the form of GO annotations, that it proves that the association rules produced by our system are good and this may be referred by any future research in this area[28].

The usage of XML data in the World Wide Web and elsewhere as a standard for the exchange of data and to represent semi structured data tends to develop the various tools and techniques to perform various data mining operations on XML documents and XML repositories. In recent years, several encouraging methods have been identified and developed for mining XML data. Presented an improved framework for mining association rules from XML data using XQUERY and. NET based implementation of Apriori algorithm[30].

Jochen Hipp *et al* provided several efficient algorithms that cope up with the popular and computationally expensive task of association rule mining with a comparison of these algorithms concerning efficiency[13]. He proposed that the algorithms show quite similar runtime behavior in their experiments. Komal Khurana and Mrs. Simple Sharma presented a paper[41]. This[41] paper represents comparison of five association rule mining algorithms: AIS, SETM, Apriori, AprioriTID and Apriori Hybrid. The AprioriTID and Apriori Hybrid have been proposed to solve the problem of Apriori algorithm. From the comparison they conclude that the Apriori Hybrid is better than Apriori and AprioriTID because it reduces overall speed and improves the accuracy. Ziauddin *et al* researched on

association rule mining. They presented a survey of research work since its beginning[39]. He however proposed that association rule mining is still in a stage of exploration and development. There are still some essential issues that need to be studied for identifying useful association rules.

Pratima Gautam presented an efficient version of Apriori algorithm for mining multi-level association rules in large databases to finding maximum frequent itemset at lower level of abstraction. They proposed a new, fast and an efficient algorithm (SC-BF Multilevel) with single scan of database for mining complete frequent itemsets. The proposed algorithm can derive the multiple-level association rules under different supports in simple and effective way[26].

In 1995 Houtsma presented SETM (SET-oriented Mining of association rules)[25] was motivated by the desire to use SQL to compute large itemsets. It utilized only simple database primitives, viz. sorting and merge-scan join. It was easy, rapid and durable over the variety of parameter values. It proved that some aspects of data mining can be carried out by using general query languages such as SQL, instead of developing specialized black-box algorithms. The set-oriented feature of SETM eased the development of extensions Apriori.

In 1997 Cheung presented A conditional FP-tree is in orders of magnitude smaller rivaled to the global FP-tree. Consequently the size of the FP-trees to be handled would be considerably dwindled when a conditional FP-tree is created out of each projected database. This has been proved to be quicker than the Tree-Projection algorithm[7] where in the database is projected recursively into a tree of databases.

In 2012 Sanjeev Rao, Priyanka Gupta[31] proposed a novel scheme for mining association rules pondering the number of database scans, memory consumption, the time and the interestingness of the rules. They removed the disadvantages of APRIORI algorithm by determining a FIS data extracting association algorithm which is proficient in terms of number of database scan and time. They eradicate the expensive step candidate generation and also avoid skimming the database over and again. Thus they used Frequent Pattern (FP) Growth ARM algorithm that is more effectual structure to extract patterns when database intensifies.

In 2008 Kamrul *et al*[21] presented a novel algorithm Reverse Apriori Frequent pattern mining, which is a new methodology for frequent pattern production of association rule mining. This algorithm works proficiently, when the numerous items in the enormous frequent itemsets is near to the number of total attributes in the dataset, or if number of items in the hefty frequent itemsets is predetermined.

Association Rule Mining Algorithms

Association rule problem

Let $I= I_1, I_2, \ldots\ldots, I_m$ being a set of m different qualities, T be transacted that comprises a set of items such that $T \subseteq I$, D, D be a database with altered contract records Ts. An association rule is a suggestion in the procedure of $X \cap Y$, where X, $Y \subset I$ are arrays of items named itemsets, and $X \cap Y = \phi$. X is named designer though Y is called consequential, the rule capitals X indicates Y. There are two significant simple processes for association rules, support (s) and confidence (c). Then the database is huge and users' concern almost only individuals commonly obtained items, commonly beginnings of support and assurance are predefined by operators to drop those rules that are not so exciting or useful. The two edges are called minimal support and minimal confidence individually, supplementary limits of stimulating rules also can be identified by the operators. The two elementary parameters of Association Rule Mining are maintenance and

declaration. Support (s) of an association rule is definite as the fraction of records that comprise XUY to the entire integer of records in the database. The computation for each item is improved by one all time the item is faced in different operation T in database D during the scanning process. It means the support count does not take the magnitude of the item into account. For instance, in a transaction a customer purchase three cups of a tea, but we only increase the support count number of tea one by one, in another term if a transaction contains an item then the support count of this item is increased by one. Support (s) is premeditated by the succeeding.

$$\text{Support (XY)} = \frac{\text{Support total of XY}}{\text{Total digit of Operation during D}}$$

Support is used to seek out the strongest association rules within the item sets.

Confidence is another approach for locating the association rules. Confidence of AN association rule is outlined because the percentage/fraction of range the amount the quantity of transactions that contain X Y to the entire number of records that contain X, wherever if the proportion exceeds the edge of confidence a motivating association rule X=>Y will be generated.

$$\text{Confidence (X|Y)} = \frac{\text{Support (XY)}}{\text{Support (X)}}$$

Positive association rules

The normal convention in discovering the association rules are by suggests that of any frequent item sets that area unit gift within the given transactional information. the principles that area unit ordinarily obtained by suggests that of mistreatment minimum support threshold and minimum confidence threshold area unit usually referred because the positive association rules and therefore the rule is of the shape ¬A ¬B. which means that they're capable of associating one component to the opposite component in a very given set of transactional records.

Negative association rules

Contrary to the positive association rules delineated on top of, negative association rules area unit defmed because the rule that involves the absence of item sets. For instance, contemplate A => ¬B, here, "¬", indicates the absence of Associate in Nursing item set B in a very set of given transactional records. The foundations of the forms (A→ ¬B, ¬A→ B and ¬A→¬B) area unit negative association rules[33]. Constraints based association rule mining: In associate degree interactive mining setting, it becomes a necessity to modify the user to precise his interests through constraints on the discovered rules, and to alter these interests interactively. The most notable constraints are item constraints, that are those who impose restrictions on the presence or absence of things in an exceedingly rule. These constraints may be within the variety of conjunction or a disjunction. Such constraints are introduced initial in wherever a replacement methodology, for incorporating the constraints into the candidate generation section of the Apriori formula, was projected. During this manner, candidates are assured to adjust the Item constraints besides the initial support and confidence constraints. They outlined what's referred to as affected Frequent Queries (CAQs) and bestowed an excellent classification of constraints constructs that may be exploited in them by introducing the notions of concise and anti-monotone constraints. The CAP (Constrained Apriori) was presented for economical discovery of affected association rules[2].

Apriori algorithm

Apriori Algorithm was initially introduced by R. Agrawal. In this algorithm is used for frequent item set, Association rule mining techniques. Apriori uses pruning

techniques to avoid measure bound item sets, whereas guaranteeing completeness[1]. The Apriori algorithmic program is predicated on the Apriori principle[14]. The algorithm for use a level-wise search, k-itemsets exist used to explore (k+1)-itemsets, to mine frequent itemsets from transactional database for Boolean association rules. In this algorithm, frequent subsets are extended one item at a time and this step is known as candidate generation process. Then groups of candidates are tested against the data. To count candidate item sets efficiently, Apriori uses breadth-first search method and a hash tree structure.

There are several key concepts used in Apriori algorithm such as Frequent Itemsets, Apriori Property and Join Operation. (See figure 2.)

It identifies the frequent individual things within the information and extends them to larger and bigger item sets as long as those item sets seem sufficiently typically within the information. Apriori algorithmic rule confirms frequent item sets that may be accustomed determine association rules that highlight general trends within the information.

Procedure for Apriori algorithm

1. CIk: Candidate itemset having size k
2. FIk: Frequent itemset having size k
3. FI1 = Frequent itemset;
4. For (k=1; FIk! = null; k++) do
5. BeginCIk+1 = candidates generated from FIk;
6. For each dealing t in info D do
7. Increment the count price of all candidates in
8. CIk+1 that area unit contained in t
9. FIk+1 = candidates in CIk+1 with min_support
10. End
11. Return FIk (See figure 3.)

The table 1 shows the performance of Apriori Algorithm based on data support, speed and accuracy and the same is depicted on bar chart in figure 4.

AprioriTid algorithm

AprioriTID algorithmic rule uses the generation operate so as to work out the candidate item sets. The sole distinction between the two algorithms is that, in AprioriTID algorithmic rule the info isn't referred for investigating support once the primary pass itself. Here a group of candidate item sets is employed for this purpose for k>1. Once a group action doesn't have a candidate k-item set in such a case the set of candidate item sets won't have any entry for that group action. This can decrease the quantity of group action within the set containing the candidate item sets Compared to the information. As worth of k will increase each entry can become smaller than the corresponding transactions because the variety of candidates within the transactions can persevere decreasing. Apriori solely performs higher than AprioriTID within the initial passes however a lot of passes area unit given AprioriTID definitely has higher performance than Apriori. Database isn't used for count the support of candidate itemsets once the primary pass. The method of candidate itemset generation is same just like the Apriori rule. Another set C' is generated of that every member has the TID of every dealing and therefore the massive itemsets gift during this dealing. The set generated i.e. C' is employed to count the support of every candidate itemset. (See figure 5.)

The table 2 shows the performance of AprioriTid Algorithm based on data support, speed and accuracy and the same is depicted on bar chart in figure 6.

Apriori hybrid algorithm

Apriori Hybrid algorithm was initially introduced by R. Agrawal in 1994. Apriori

and AprioriTID use constant candidate generation procedure and computation constant item sets. Apriori examines each dealing within the information. On the opposite hand, instead of scanning the information, AprioriTID scans candidate item sets utilized in the previous pass for getting support counts. Apriori Hybrid uses Apriori within the initial passes and switches to AprioriTid once it expects that the candidate item sets at the tip of the pass are going to be in memory. As Apriori will higher than AprioriTid within the earlier passes and AprioriTid will higher than Apriori within the later passes.

The table 3 shows the performance of Apriori hybrid Algorithm based on data support, speed and accuracy and the same is depicted on bar chart in figure 7.

AIS Algorithm

The AIS [Artificial immune system] algorithm was the first algorithm proposed by Agrawal[2], this algorithmic program is used to sight frequent item sets. It uses candidate generation so as to sight them. The candidate's area unit generated on the fly and that they area unit then compared with the already generated frequent item sets. One in every of the disadvantage of this algorithmic program contains the generation and tally of too several candidate item sets that prove to be little. This was the primary algorithmic program to introduce the matter of generation of association rules. (See figure 8.)

The table 4shows the performance of AIS Algorithm based on data support, speed and accuracy and the same is depicted on bar chart in figure 9.

SETM Algorithm

In the SETM algorithmic rule, candidate itemsets square measure generated on-the-fly because the information is scanned, however counted at the top of the pass. Then new candidate itemsets square measure generated a similar means as in AIS algorithmic rule, but the transaction symbol TID of the generating group action is saved with the candidate itemset in a very sequent structure. It separates candidate generation method from reckoning. At the end of the pass, the support count of candidate itemsets is determined by aggregating the sequent structure. The SETM algorithmic rule has a similar disadvantage of the AIS algorithmic rule. Another disadvantage is that for every candidate itemset, there square measure as several entries as its support value[2]. (See figure 10.)

The table 5shows the performance of AIS Algorithm based on data support, speed and accuracy and the same is depicted on bar chart in figure 11.

FP-Growth algorithm (Frequent pattern)

The FP-Growth Algorithm, proposed by J. Han. In FP-growth needs constructing FP-tree. Is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for storing compressed and crucial information about frequent patterns named frequent-pattern tree (FP-tree)[18]. For that, it needs 2 passes. FP growth uses divide and conquer strategy. It needs 2 scans on the info. It 1st computes a listing of frequent items sorted by frequency in dropping order (F-List) and during its 1st info scan. Within the second scan, the database is compressed into a FP-tree. This algorithmic rule performs mining on FP-tree recursively. There's a tangle of finding frequent itemsets that is regenerate to looking and constructing trees recursively. The frequent itemsets area unit generated with solely 2 passes over the info and without any candidate generation method. There area unit 2 sub methods of frequent patterns generation process that includes: construction of the FP-tree, and generation of the frequent patterns from the FP-tree. (See figure 12.)

FP-tree is made over the data-set exploitation a pair of passes are as follows:

Pass 1

- Scan the info and realize support for every item.
- Discard rare things.
- Type frequent things in downward order that is based on their support.
- By exploitation this order we will build FP-tree, so common Prefixes will be shared.

Pass 2

- Here nodes correspond to things and it's a counter.
- FP-growth reads one dealings at a time then maps it to a path.
- Mounted order is employed, so methods will overlap once transactions share the things.

In this case, counters are incremented. Some pointers are maintained between nodes that contain identical item, by creating on an individual basis coupled lists. The lot of methods that overlap, higher the compression. FP-tree could slot in memory. Finally, frequent itemsets are extracted from the FP-Tree.

The Procedure FP-growth (Tree T, A)

If Tree T contains a single path P, Then for each combination of the nodes in the path P do Generate pattern B U A with support = minimum support of nodes in B

Else for each Hi in the header of the tree T do

{Generate pattern B=Hi U A with support = Hi. support;

Construct B's conditional pattern base and B's conditional FP-
Tree that is B;
If Tree B ≠Ø
Then call FP-growth (Tree B, B)}
FP-tree Example

Step 1

By-Product of First Scan of Database. (See table 6.)

Scan DB for the first time to generate L. (See table 7.)

Step 2

Scan the dB for the second time, order frequent things in every dealing. (See table 8.)

The table 9shows the performance of AIS Algorithm based on data support, speed and accuracy and the same is depicted on bar chart in figure 13.

COMPARISON

The comparative study of six algorithms are shown in the Table 10, algorithmic aspects of association rule mining are dealt with. From an extensive variation of efficient algorithms the most important ones are compared. The algorithms are systemized and their performance is analyzed based on runtime and theoretical considerations. Despite the identified fundamental differences concerning employed strategies, runtime shown by algorithms is almost similar. FP growth displayed better performance in all the cases leaving Apriori behind by making only 2 passes to the data sets and abolishing the concept of candidate generation. The paper would give a basic idea to the company's data mining team about the algorithm which would yield better results.

The demerits and merits of the six ARM algorithms are shown in table 11. Based on the literature survey, the six algorithms are compared using data support, speed in initial phase, speed in later phase and accuracy measurements. In data support aspect, AIS, SETM work well on small database Apriori work good for medium size data bases, and AprioriTid, Apriori hybrid and Fp-growth well suited for large data bases. In speed in initial phase AIS, SETM and AprioriTID work slow speed in the first phase, Apriori, Apriori hybrid, Fp-growth

work well on fast speed in the starting phase. In speed in later phase AIS, SETM and Apriori work slow speed in final phase. AprioriTID, Apriori hybrid and FP-growth well work on fast speed in final phase. In accuracy, AIS is very less accurate. SETM and Apriori are less accurate measurement. AprioriTID are medium accurate measurement, it's more accurate than Apriori. Apriori hybrid are high accurate, and it's more accurate than AprioriTID. FP-growth work well on very high accurate in all six algorithms. The Figure 14 show the overall comparison between association rule mining algorithms and its performance.

CONCLUSION

This paper presents the extensive of study of various ARM algorithms in data mining which are really useful and very much needed to obtain useful facts or associations among data items in large data sets to take some important decision making in any kind of problems. This paper gives the overview of six ARM algorithms namely AIS, SETM, AprioriTid, Apriori hybrid and FP-growth in which all algorithms are analyzed and the merits and demerits are reported. In comparative study, all six algorithms has been compared with respect to three important criteria such as Data Support, speed and accuracy. Based on speed, the Apriori hybrid algorithm equally good as FP-Growth. However, the FP-Growth algorithm outperforms well than the Apriori hybrid with respect to Accuracy. The comparative result shows that the FP-Growth algorithm is more suitable for obtaining significant associations from very large datasets in a speedy and accurate manner.

REFERENCES

1. Agrawal, R. and Srikant, R. 1995. Mining sequential patterns, P. S. Yu and A. S. P. Chen, Eds. IEEE Computer Society Press, Taipei, Taiwan, 3, 14.
2. Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. Proceedings of the international conference on very large data bases (pp. 407–419).
3. Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association between sets of items in massive database. International proceedings of the ACM-SIGMOD international conference on management of data (pp. 207–216).
4. Attila Gyenesei, A Fuzzy approach for mining quantitative association rules, Technical Report: TUCS-TR-336 Year of Publication, 2000.
5. Borgelt, C. "Efficient Implementations of Apriori and Eclat". Workshop of frequent item set mining implementations (FIMI 2003, Melbourne, FL, USA).
6. Brijs, T., Goethals, B., Swinnen, G., Vanhoof, K., & Wets, G. (2000). A data mining framework for optimal product selection in retail supermarket data: The generalized PROFSET model. Proceedings of the ACM-SIGKDD international conference on knowledge discovery and data mining (pp. 20–23).
7. D.W.L. Cheung, S.D. Lee, and B. Kao., 1997. "A general incremental technique for maintaining discovered association rules" In Database Systems for advanced Applications, pp. 185-194.
8. Dong, G., & Li, J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. Proceedings of the international conference on knowledge discovery and data mining (KDD'99) (pp. 43–52).
9. Fayyad, U., Djorgovski, S. G., & Weir, N. (1996). Automating the analysis and cataloging of sky surveys. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining (pp. 471–494). Cambridge, MA: MIT Press.
10. Ha, S., Bae, S., & Park, S. (2000). Web mining for distance education. In IEEE international conference on management of innovation and technology (pp. 715–719).

11. Han, J. and Kamber, M. 2000. Data Mining Concepts and Techniques. Morgan Kanufmann.
12. Hilderman, R. J., & Hamilton, H. J. (2001). Evaluation of interestingness measures for ranking discovered knowledge. Proceedings of the Pacific-Asia conference on knowledge discovery and data mining (PAKDD'01) (pp. 247–259).
13. Hipp, J., G¨untzer, U., and Nakhaeizadeh, G. "Algorithms for Association Rule Mining – A General Survey and Comparison", SIGKDD Explorations ACM, JULY 2000.
14. http://www.users.cs.umn.edu/~kumar/dmbook/ch6.pdf. Association analysis Basic concepts and algorithms.
15. HUANG Liusheng, CHEN Huaping, WANG Xun, CHEN Guoliang, a Fast Algorithm for Mining Association Rules, *J. Comput. Sci. & Technol*., Vol. 15 No. 6, pp 619-624, Nov. 2000.
16. Hunyadi, D. "Performance comparison of Apriori and FP-Growth Algorithms in Generating Association Rules". Proceedings of the European Computing Conference ISBN: 978-960-474-297-4.
17. Irina Tudor, Association rule mining as a data mining technique, BULETINUL universitatii Petrol-Gaze din Ploiesti, Vol. LX No1/ 2008, page 49-56.
18. J. Han, H. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. In: Proc. Conf. on the Management of Data (SIGMOD'00, Dallas, TX). ACM Press, New York, NY, USA 2000.
19. J. Han, J. Pei, and Y. Yin, Miniing Frequent Patterns without Candidate Generation, Proceedings of th e ACM SIGMOD, Dallas, TX, May 2000, pp. 1-12.
20. J. S. Park, M.-S. Chen, and P. S. Yu, An effective Hash-Based Algorithm for Mining Association Rules, Proceedings of the ACM SIGMOD, San Jose, CA, May1995, pp. 175-186.
21. Kamrul, Shah, Mohammad, Khandakar, Hasnain, Abu, 2008. "Reverse Apriori Algorithm for Frequent Pattern Mining", Asian Journal of Information Technology, pp.524-530, ISSN: 1682-3915.
22. Kiran R. U., and Reddy P. K.: An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules. http://www.iiit.net/techreports/2009_24.pdf.
23. Kitts, B., Freed, D., &Vrieze, M. (2000). Cross-sell: A fast promotion tunable customer-item recommendation method based on conditionally independent probabilities. Proceedings of the ACM-SIGMOD conference on knowledge discovery and data mining (KDD'00) (pp. 437–446).
24. LI Pingxiang, CHEN Jiangping, BIAN Fuling, A Developed Algorithm of Apriori Based on Association Analysis, *Geo-spatial Information Science*, Vol. 7, Issue 2, pp 108-112, june 2004.
25. M. Houtsma, and Arun Swami, 1995. "Set-Oriented Mining for Association Rules in Relational Databases". IEEE International Conference on Data Engineering, pp. 25–33.
26. Pratima Gautam and K.R. Pardasani, "Algorithm for Efficient Multilevel Association Rule Mining" *In (IJCSE) International Journal on Computer Science and Engineering*, Volume 02, No. 05, 1700-1704, 2010.
27. Qiankun Zhao, Sourav S. Bhowmick, and Association Rule Mining: A Survey, Technical Report, CAIS, Nanyang Technological University, Singapore, 2003.
28. R. Porkodi, Dr. B.L Shivakumar finding relationships among gene ontology terms in biological documents using Association Rule mining and GO annotations ISSN: 2249-9555 Vol. 2, No.3, June 2012, pp. 542-550.
29. R. Porkodi, Dr. B.L Shivakumar Rule based approach for constructing Gene/Protein names Dictionary from Medline abstract, ISSN 2277–9140, July 2012, p.p457-468.
30. R. Porkodi, Dr. B.L Shivakumar, An Improved Association Rule Mining Technique for Xml Data Using Xquery and Apriori Algorithm, pp.1510 – 1514, March 2009.
31. S. Rao, P. Gupta 2012. "Implementing improved algorithm over Apriori data mining association rule algorithm", *IJCST*, vol. 3, pp.489-493, ISSN: 2229-4333.
32. Song, H. S., Kim, J. K., & Kim, S. H. (2001). Mining the change of customer behavior in an internet shopping mall. *Expert Systems with Applications*, 21,158–168.

33. Srikant, R. & Agrawal, R., "Mining quantitative association rules in large relational tables", SfGMOD Rec., ACM, 1996, 25.
34. Tan, P. N., &Kumar, V. (2000). Interestingness measures for association patterns: A perspective. KDD 2000 Workshop on Post processing in Machine Learning and Data Mining, Boston, MA, August.
35. Virendra kumar Shrivastava, Dr. parveen kumar and DR. K.R. pardasani, FP-Tree and COFI Based App roach for Mining of Multiple Level Association Rules in Large database, IJCSIS, *International Journal of Computer Science and Information Security*, Vol.7 No. 2, 2010.
36. Wang, K., Zhou, S., & Han, J. (2002). Profit mining: From patterns to actions. Proceedings of international conference on extending data base technology (EDBT'02) (pp. 70–87).
37. Xindong Wu and *et al*, Top 10 algorithms in data mining, KnowlInfSyst, Springer-Verlag London Limited, pp 14:1–37, 2008.
38. Zhang, S., Zhang, C., & Yan, X. (2003). Post-mining: Maintenance of association rules by weighting. Information Systems, 28, 691–707.
39. Ziauddin, Kammal, S., Khan, Z.K., and Khan, I.M., "*Research on association rule mining*". ACMA Volume 2, no. 1, 2012, ISSN 2167-6356. World science Publisher, United States, 2012.
40. Zhu Ming, datamining, University of Science and Technology, china Press, Hefel, pp: 115 – 126, 2002.
41. Khurana, K., and Sharma, S. "A comparative analysis of association rule mining algorithms". *International Journal of Scientific and Research Publications*, Volume 3, Issue 5, May 2013.

Table 1. Performance of the Apriori algorithm

Features	Apriori
Data support	Limited
Speed in initial phase	High
Speed in later phase	Slow
Accuracy	Less

Table 2. Performance of the AprioriTid algorithm

Features	AprioriTID
Data support	Often suppose large
Speed in initial phase	Slow
Speed in later phase	High
Accuracy	More accurate than Apriori

Table 3. Performance of the Apriori hybrid algorithm

Features	Apriori hybrid
Data support	Very Large
Speed in initial phase	High
Speed in later phase	High
Accuracy	More accurate than AprioriTID

Table 4. Performance of the AIS algorithm

Features	AIS
Data support	Less
Speed in initial phase	Slow
Speed in later phase	Slow
Accuracy	Very less

Table 5. Performance of the SETM algorithm

Features	SETM
Data support	Less
Speed in initial phase	Slow
Speed in later phase	Slow
Accuracy	Less

Table 6. First scan of database

Item	frequency
f	4
c	4
a	3
b	3
m	3

Table 7. Itemsets for first time to generate L

TID	Items bought
100	{f, a, c, d, g, i, m, p}
200	{a, b, c, f, l, m, o}
300	{b, f, h, j, o}
400	{b, c, k, s, p}
500	{a, f, c, e, l, p, m, n}

Table 8. Database second time order frequent

TID	Items bought	frequent items
100	{f, a, c, d, g, l, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

Table 9. Performance of the FP-Growth algorithm

Features	FP-growth
Data support	Very Large
Speed in initial phase	High
Speed in later phase	High
Accuracy	More Accurate

Table 10. Comparative study of algorithms

S. No.	Algorithms	Algorithm Data support	Merits	Demerits	Year
1.	Apriori[5]	Best used for closed item sets.	Fast Less candidate sets. Generates candidate sets from only those items that were found large.	Takes a lot of memory.	2003
2.	AprioriTID[2]	Used for minor itemsets.	Better than SETM. Better than Apriori for small databases, Time saving.	Doesn't use whole database to count candidate sets.	1994
3.	SETM[2]	Not frequently used.	Separates generation from counting.	Very large execution time. Size of candidate set large.	1994
4.	Apriori Hybrid[2]	Used where Apriori and AprioriTID used.	Better than both Apriori and AprioriTID.	An extra cost is incurred when shifting from Apriori to AprioriTid	1994
5.	AIS[2]	Not frequently used, but when used is used for small itemsets.	Better than SETM. Easy to use	Candidate sets generated on the fly. Size of candidate set large.	1994
6.	FP-Growth[16,5]	Used in cases of large itemsets as it doesn't require generation of candidate sets.	Only 2 passes of dataset. Compresses data set. No candidate set generation required so better than éclat, Apriori.	Using tree structure creates complexity.	2003

Table 11. Comparison of association rule mining algorithms

Features	AIS	SETM	Apriori	AprioriTID	Apriori hybrid	FP-growth
Data support	Less	Less	medium	Often suppose large	Very Large	Very Large
Speed in initial phase	Slow	Slow	High	Slow	High	High
Speed in later phase	Slow	Slow	Slow	High	High	High
Accuracy	Very less	Less	Less	Medium, More accurate than Apriori	Fast, More accurate than AprioriTID	Very fast, More Accurate

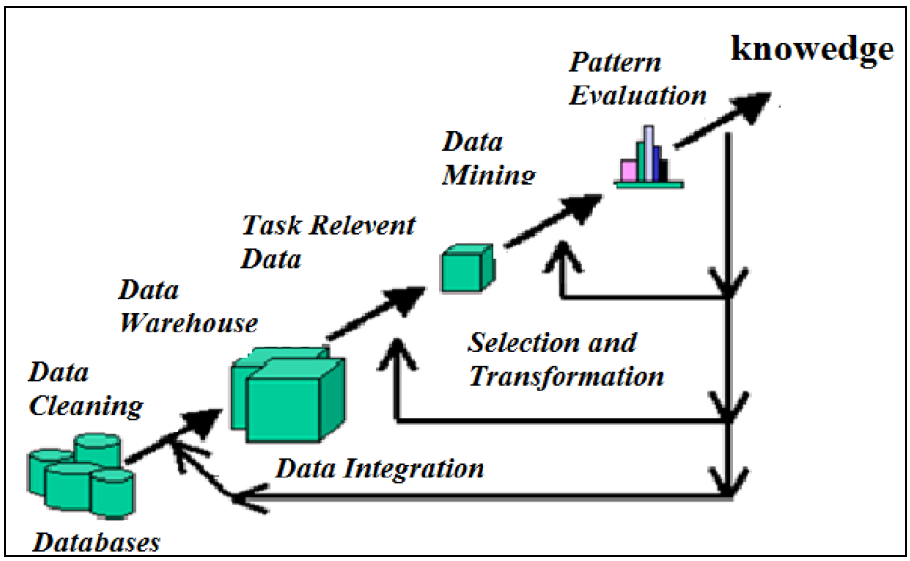

Figure 1. Knowledge Discovery in Database processes

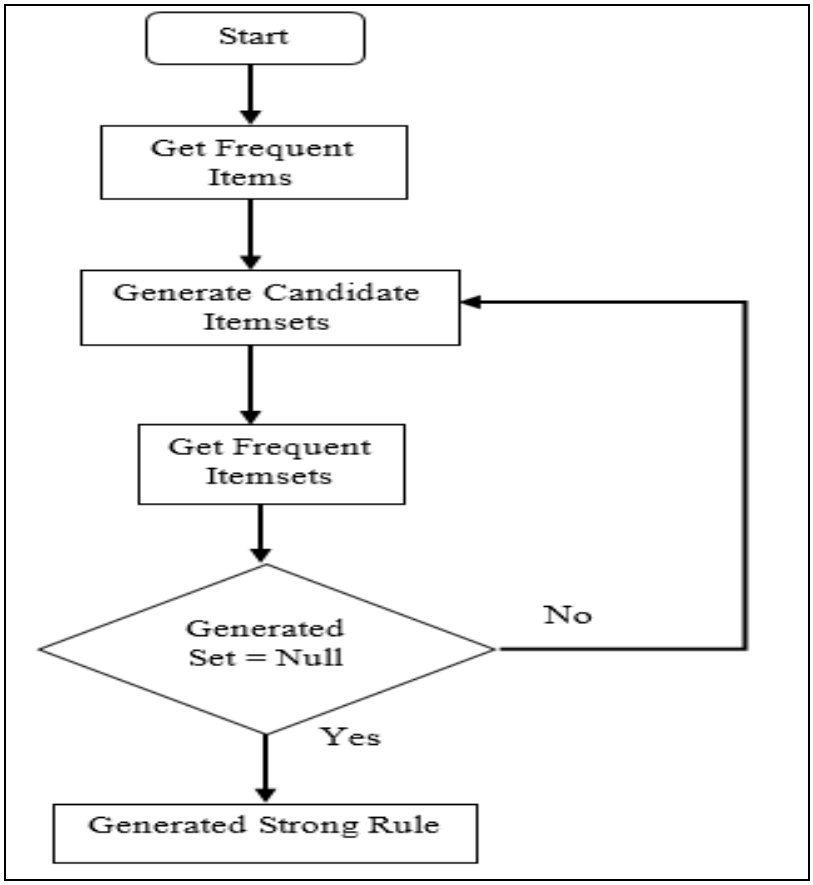

Figure 2. Apriori algorithm process

Items	Count Number
I1	7
I2	8
I3	6
I4	2
I5	3
I6	1

a) C1

Large 1 Items
I1
I2
I3
I5

b) L1

Items	Count Number
I1,I2	5
I1,I3	4
I1,I5	3
I2,I3	4
I2,I5	3
I3,I5	1

c) C2

Large 2 Items
I1,I2
I1,I5
I2,I5
I2,I3
I1,I3

d) L2

Items	Count Number
I1,I2,I5	3
I1,I2,I3	2

e) C3

Figure 3. Sample Itemsets in Apriori algorithm

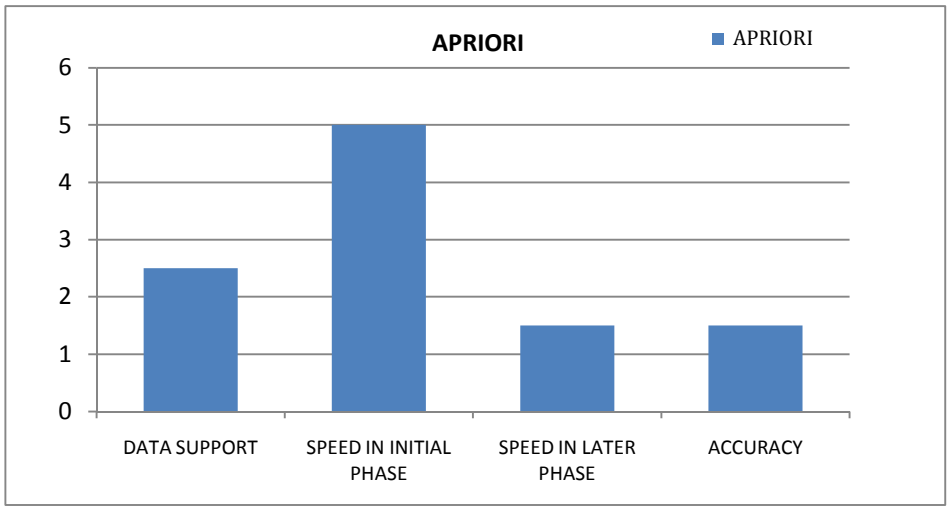

Figure 4. Graphical representation of the performance of Apriori algorithm

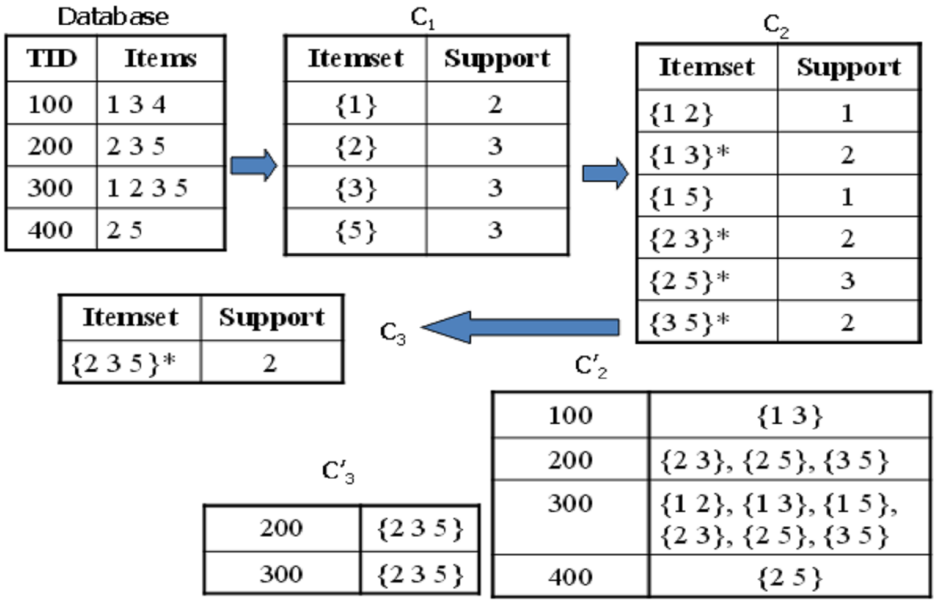

Figure 5. Sample itemsets in AprioriTid algorithm

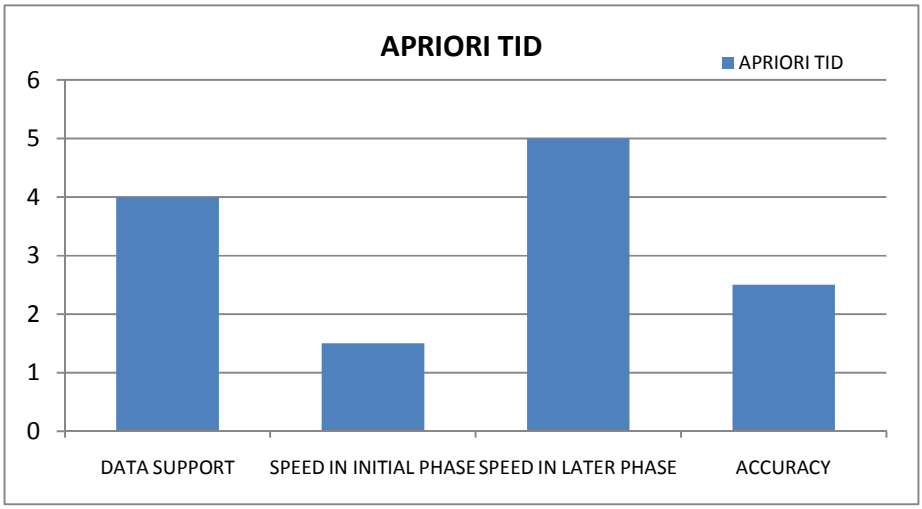

Figure 6. Graphical representation of the performance of Apriori TID algorithm

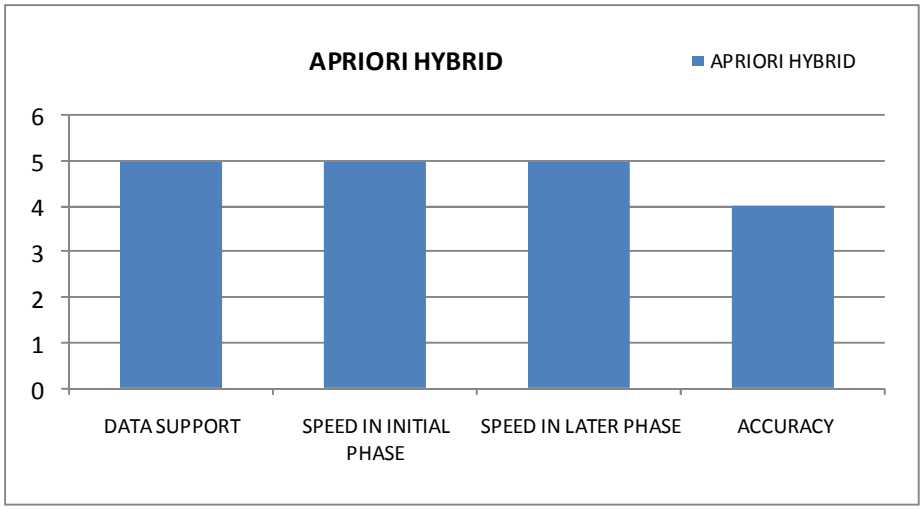

Figure 7. Graphical representation of the performance of Apriori hybrid algorithm

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	Support
{1 3}*	2
{1 4}	1
{3 4}	1
{2 3}*	2
{2 5}*	3
{3 5}*	2
{1 2}	1
{1 5}	1

C_3

Itemset	Support
{1 3 4}	1
{2 3 5}*	2
{1 3 5}	1

Figure 8. Sample itemsets in AIS Algorithm

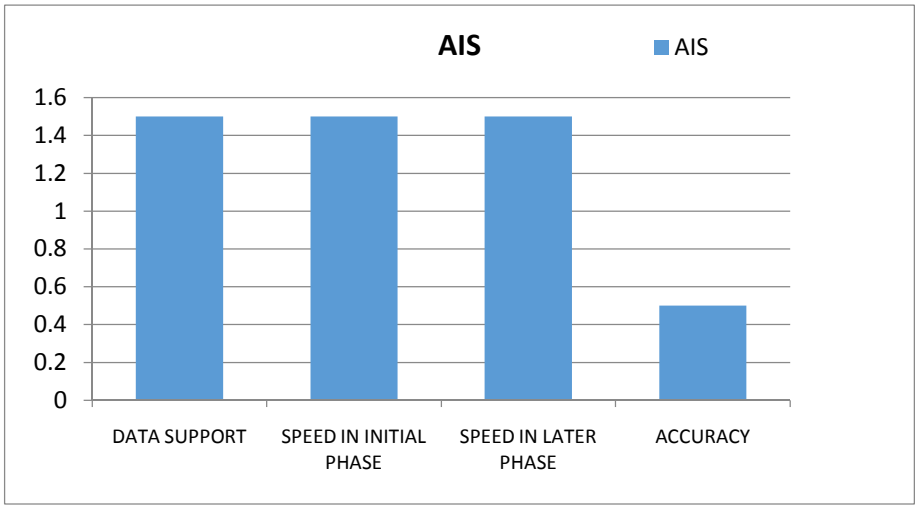

Figure 9. Graphical representation of the performance of AIS algorithm

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	TID
{1 3}	100
{1 4}	100
{3 4}	100
{2 3}	200
{2 5}	200
{3 5}	200
{1 2}	300
{1 3}	300
{1 5}	300
{2 3}	300
{2 5}	300
{3 5}	300
{2 5}	400

C_3

Itemset	TID
{1 3 4}	100
{2 3 5}	200
{1 3 5}	300
{2 3 5}	300

Figure 10. Sample Itemsets in SETM Algorithm

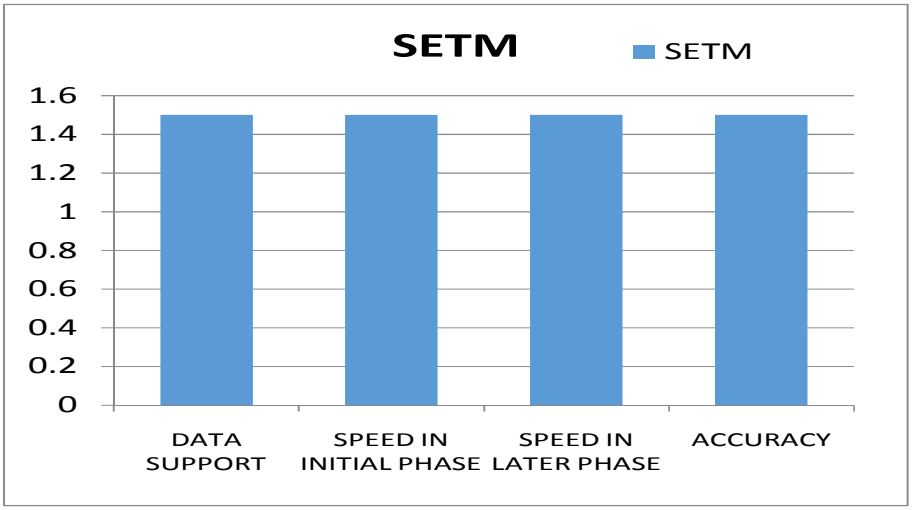

Figure 11. Graphical representation of the performance of SETM algorithm

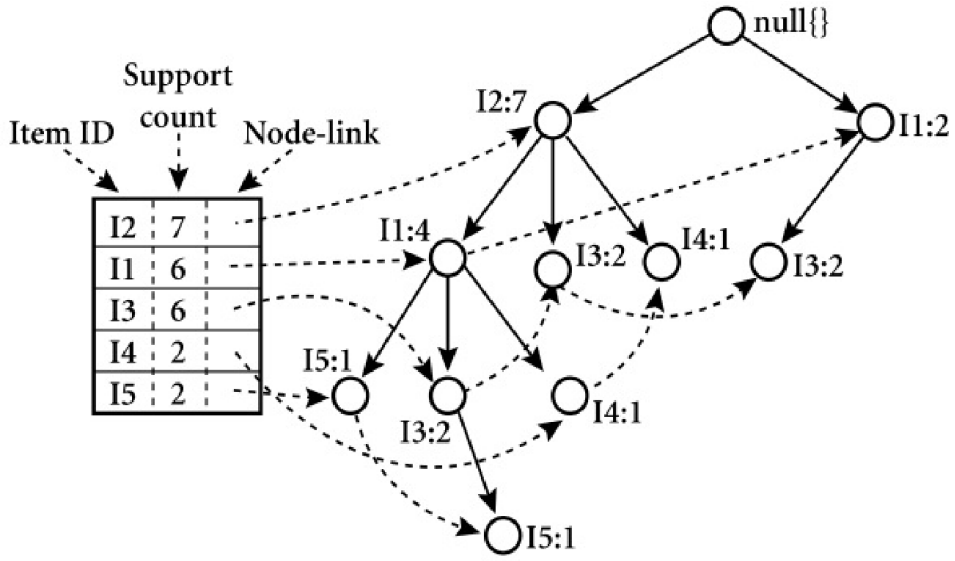

Figure 12. Sample for FP-Tree

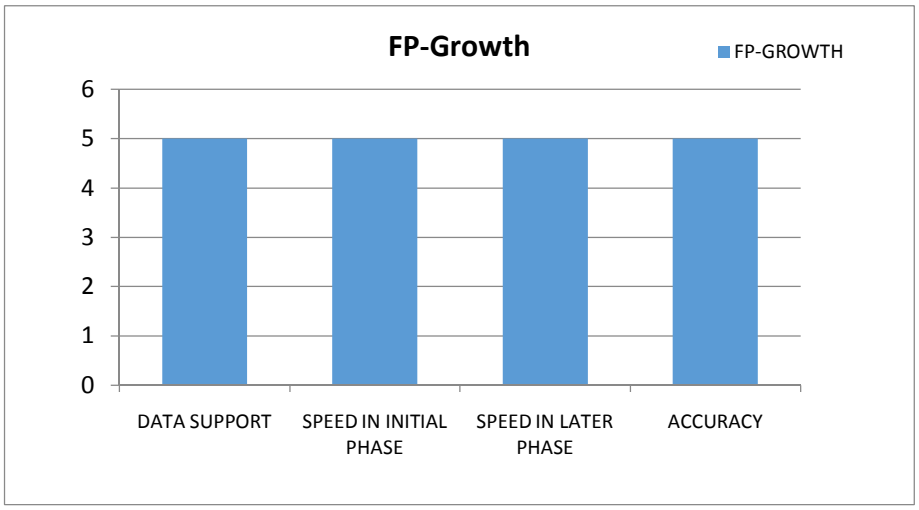

Figure 13. Graphical representation of the performance of FP-Growth algorithm

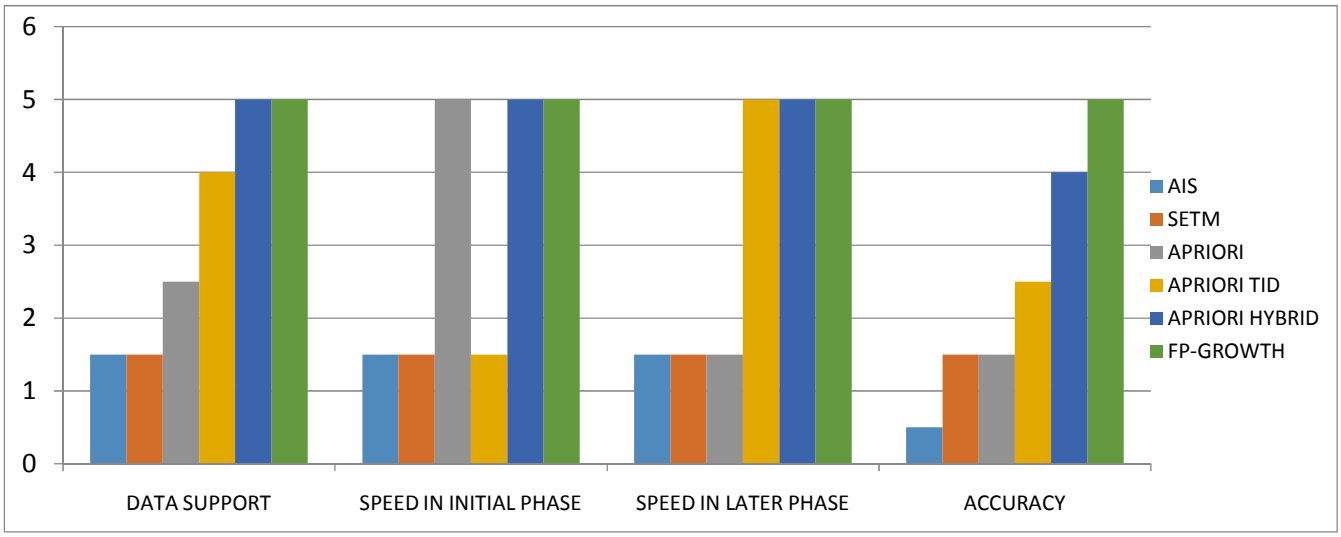

Figure 14. Comparisons of association rule mining algorithms