iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** XII **Month of publication:** December 2020

DOI: https://doi.org/10.22214/ijraset.2020.32589

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue XII Dec 2020- Available at www.ijraset.com

Frequent Itemsets on Temporal Data by using Apriori and FP-Growth Algorithm

Rupali Bukte[1], Mr. Sunil Jaiswal[2]
[1]*M-Tech ll[nd] Year, MGM's Jawaharlal Neharu Engineering College, Aurangabad*
[2]*Assistant Professor, Computer science and Engineering, MGM's Jawaharlal Neharu Engineering College, Aurangabad*

Abstract: ***FP-Growth is an association rule mining which dependent on finding frequent itemsets on temporal data sets, which doesn't have to create a candidate generation key. The data mining uses different methodology to finding frequency algorithm .In this paper we are using a both algorithms i.e Apriori algorithm and FP-Growth algorithm. The Apriori algorithm needs a multiple scans which consumes more memory and more time, and Apriori needs a candidate generation, to over come this problem we are implementing the fp-growth algorithm which does not need a candidate generation key, and it is a more efficient and more accurate than apriori. In this paper we are trying to increase efficiency and accuracy of the algorithm in less time and to show which algorithm is more better.***
Keywords: Data mining, Association mining rule, Apriori , FP-growth, frequent itemsets .

I. INTRODUCTION

Data mining is used for mining the data from large amount of datasets and also mining the patterns from the datasets. Frequent item sets are traced out mainly in the transactional datasets. It is first proposed by Agarwal, Imielinski, and Swami in the context of frequent item sets and association rule mining[1].
Data mining is a research area that crosses in various disciplines such as: Database, Artificial intelligence, Statistics, Information theory, Machine learning, Natural language processing, Pattern recognition, Information retrieval, Nural network, Data visualization and many more.
The data extracted from the above mentioned cognitive fields are raw data of the web. In the data mining knowledge data discovery there are three process i.e preprocessing, data mining, post processing. The preprocessing executes when the data mining techniques are applied to the data. The preprocessing contains data cleaning, data integration, data selection and data transformation. The main process of KDD is data mining in this data mining different algorithms are applied to produce knowledge.
We take a example of a transactional database of a supermarket, we may have bought two items together i.e {milk, butter} with support of 20%. That means that 20% of all truncation contains milk and butter together. Supermarket, bank, retails, departmental stores etc, are a temporal databases.
A temporal databases are a which contains time stamping information. For example milk and butter may be order together with support count of 80% all transaction in a morning 7 to 9 A.M. while their support in the whole database is 20%,in fact interesting pattern are often releted to the specific period of the time; period of time; therefore, the time during which they can be observer is important.
The rest of paper is organized as follows. In section II, we discuss two algorithms for finding frequent itemsets with our work. In section III, the proposed both algorithms is presented. In section IV, implementation of both algorithms details regarding are discussed. In section V, experimental and result analysis is discussed. In section VI, conclusion with some discussion.

II. PAGE LAYOUT

A. Association mining rule

Association mining rule is mostly used for the finding frequents patterns, associations, corelations, or a causal structure among sets of a itemsets in a transactional database.
Association mining rule consist of two main attributes i.e Support and confidence. The first sub problem plays an important role in association rules mining. The association mining rules consist of data mining techniques i.e Apriori algorithm and fP-Growth algorithm. For this the implementation of algorithms the data preprocessing is done and after the algorithm will be applied on particular dataset.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue XII Dec 2020- Available at www.ijraset.com

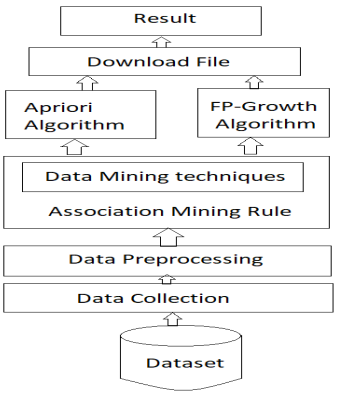


Fig-1 Association rule mining in data mining

III. NEW FRAMEWORK

Let I= {$I1$,$I2$,..In} be a set of items containing in the database T of a transaction. Each subset of I is known as an item set and each item set contains K items. Each transaction tr consists of a transaction identifier (TID), a time span ttr along with the item sets. Let T represents the total time span of the database. To find the different temporal patterns, the neighboring frequent item sets are merged. It is represented in the cubic structure as Time Cube. This Time Cube denotes the different time hierarchies. The time hierarchies are day, month, and year. The first representation of the initial cube is called Basic Time Cube (BTC).

The database is scanned for the first time; the single item patterns are generated. And it is ordered in the list L in terms of descending order of frequency. For example,{Milk:4,Butter:4,Bread:3, Curd:3,Jam:3,Peanuts:3}. This single pattern is generated according to the minimum support specified by the user. The support of the item set A is calculated as in (1),

$$\text{Support(X)} = \frac{|N(X)^{cube}|}{|N^{cube}|} \qquad (1)$$

Where $N(X)^{cube}$ represents the number of transactions containing item set A in that time interval and N^{cube} represents the number of transactions appeared in that time interval.

The avg and density of the algorithm is shown below-

$$\text{Avg} = \frac{N}{NBTCs} \qquad (2)$$

$$\text{Density} = \text{Avg} \times \propto \qquad (3)$$

Where in this the Avg represents the average transactions per basic time cube, N represents the total number of transactions and NBTC the no of basic time cube. $\propto$ is also called as the density rate.

A. Apriori Algorithm

The Apriori algorithm is one of the most basic and popular algorithm for association rule mining Agrawal and Srikanat proposed the Apriori algorithm 1994. Apriori is an algorithm which is mostly used for frequent itemset mining and association rule learning over transactional database. Apriori is designed to operate transactional database.

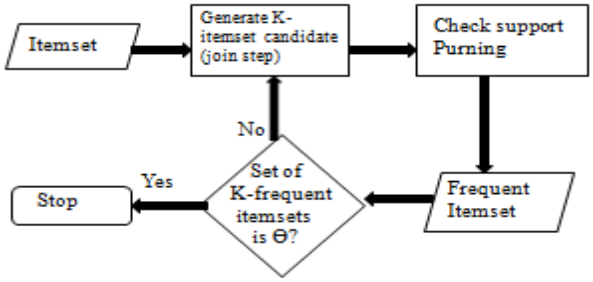


Fig-2 Flow chart for Apriori Algorithm

The Apriori algorithm uses the "bottom-up" approach, where the are determined one item at a time this is also known as candidate generation. The Apriori algorithm uses a breadth-first-search. The Apriori algorithm consist of two main steps i.e join and prune step. The new candidate set is generated in join step and if the candidate set contains the more frequent itemsets will be removed from the frequent itemsets and association mining this process is known as purne.

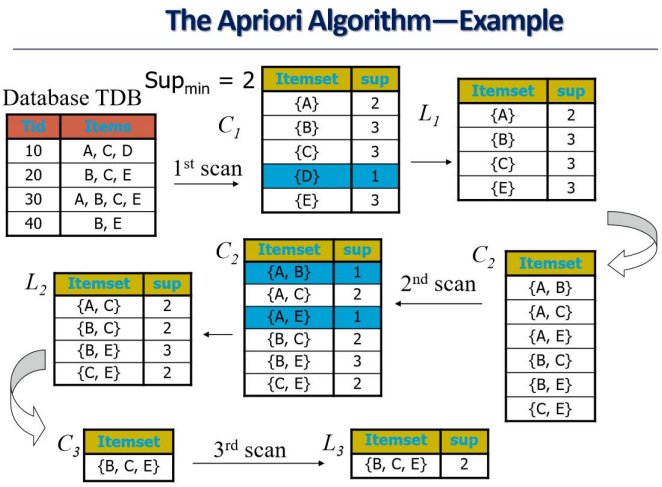


Fig-3 Example of how the Apriori Algorithm works.

The Apriori algorithm is a very simple process. It is more basic algorithm of association mining rule. It is more comfortable to learn and implementation of the algorithm.

1) Advantages of Apriori Algorithm

a) It is easy and simple to implement.
b) It uses large amount of datasets.
c) Easy to parallelized.

2) Disadvantages of Apriori Algorithm

a) The Apriori algorithm takes too much time for candidate generation.
b) The Apriori algorithm needs many scans of the database.
c) Many trivial rules are derived and it will be hard to extract the most interesting rules.
d) Rules can be inexplicable.
e) Redundent rules are generated.

B. FP-Growth Algorithm

FP-Growth algorithm was proposed by Han in 2000. FP is an improvement of a Apriori algorithm. FP-Growth is a fastest algorithm which is used for finding frequent itemsets in a data mining. In a FP-Growth algorithm a frequent pattern generated without a candidate generation. It uses a divide and conquer approach. FP-growth algorithm represents the database in the form of tree i.e known as FP tree.

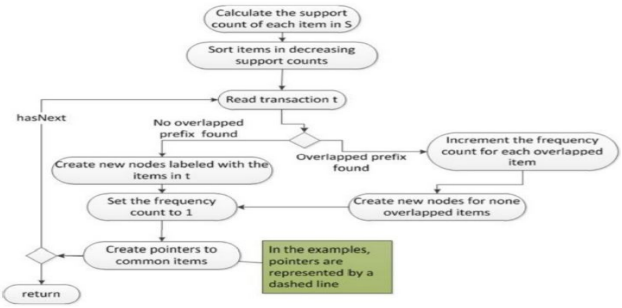


Fig-4 Flow chart for a FP-Growth algorithm

The FP-Growth algorithm is dependent on the compaction factor from the dataset at run time. FP tree is a compressed data storage structure. The FP-Growth algorithm will be performed drastically because the algorithm has to process a lot of conditional FP-Tree. The FP-Growth algorithm complexity is dependent on the path finding. FP-Growth method consist of three main stages

a) Generating conditional pattern base.
b) Generation of FP-Tress conditionals.
c) Frequent itemsets search phase.

1) Advantages of FP-Growth

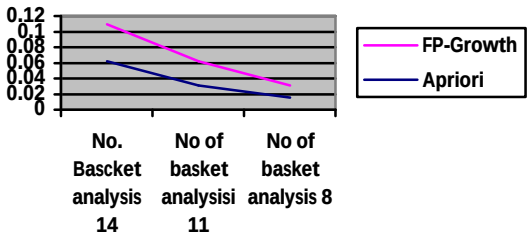
a) FP-Growth scans the database only twice.
b) FP-Growth doesn't need a candidate generation.
c) FP-Growth database is stored in a compact version of memory.
d) FP-Growth algorithm in more efficient and scalable for mining both long and short frequent patterns.

C. Comparative analysis of Apriori and FP-Growth Algorithm

Sr. No	Parameter	Apriori	FP-Growth
1.	Storage type	Array based	Tree based
2.	Search type	Breadth search	Divide and conquer
3.	Technique	Join and pure	Construct conditional frequency patterns tree which satisfy minimum support
4.	No of database scan	K+1 scan	2 scan
5.	Memory utilization	Candidate generation(large memory)	No candidate generation(less memory utilization)
6.	Database	Sparse/dense database	Large and medium dataset
7.	Run time	More time	Less time

D. Experimental results

In this section the experimental study is finished by using the two algorithms. The result will be shown on finding frequent itemsets on temporal data and comparing apriori and FP-growth algorithm.



The online retail dataset is used for this experiment, there are number of transactions. The dataset experienced with the minimum support count of 0.2 with 100 records of a transcations. The FP-growth algorithm is using a less memory and a less time. The apriori algorithm is used motre time and memory than a FP-growth.

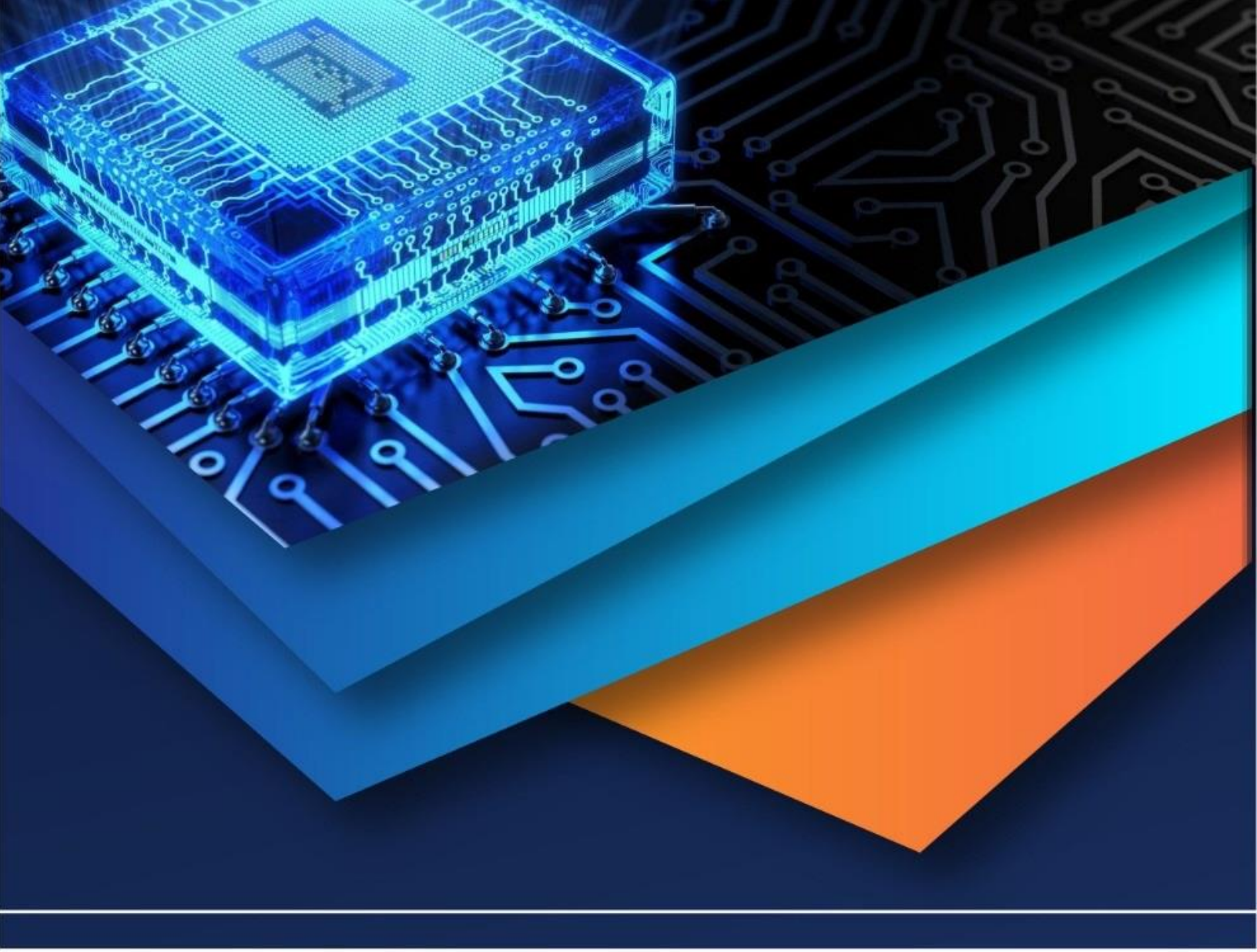
So Apriori is used for a small number of dataset and FP-growtth can be used for a large number of dataset. So FP-Growth is more better than Apriori for a large number of dataset.

IV. CONCLUSION

In this paper, we studied the mining of frequent itemsets along with temporal data by using Apriori and FP-growth algorithm. As compared to both algorithms the FP-growth algorithm gives a more accurate result and it takes less time than the apriori algorithm. The apriori algorithm gives a large utilization of a memory than a FP- growth. In FP-Growth algorithm there is no candidate generation. So for a large number of dataset the FP-growth algorithm is more better than a Apriori Algorithm. It reduces the total steps of procedure and memory utilization. From the experimental result it is concluded that the FP-growth algorithm is more better than the apriori algorithm.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Amsterdam, The Netherlands: Elsevier, 2011.
[2] R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules be- ´ tween sets of items in large databases,” ACM SIGMOD Rec., vol. 22, no. 2, pp. 207–216, 1993
[3] Borgelt, C., An Implementation of the FP-Growth Algorithm. , pp.759–770. (2005)
[4] Agarwal, R., Srikant, R.: Mining sequential patterns. In: ICDE 1995, Taipei, Taiwan, pp. 3–14, March 1995.
[5] Han, J., Kamber, M.,“Data Mining concepts and techniques”, Elsevier Inc., Second Edition, San Francisco, 2006.
[6] Y. Xiao, Y. Tian, and Q. Zhao, “Optimizing frequent time-window selection for association rules mining in a temporal database using a variable neighbourhood search,” Comput. Oper. Res., vol. 52, pp. 241–250, 2014FLEXChip Signal Processor (MC68175/D), Motorola, 1996.
[7] B. Liu, W. Hsu, and Y. Ma, “Integrating classification and association rule mining,” in Proc. 4th Int. Conf. Knowl. Discovery Data Mining, 1998, pp. 80–86.
[8] Mazaher Ghorbani and Masoud Abessi A New Methodology for Mining FrequentItemsets on Temporal Data, 0018-9391 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission
[9] Moens, Sandy, Emin Aksehirli, and Bart Goethals. "Frequent itemset mining for big data." In 2013 IEEE International Conference on Big Data, pp. 111-118. IEEE, 2013



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY