A

SEMINAR REPORT

on

"COMPARISON OF APRIORI AND FP GROWTH ALGORITHM "

Submitted in partial fulfillment of the requirement

for the assignment of Data Mining Techniques

in Computer Science and Engineering of

Jawaharlal Nehru Technological University, Kakinada

Submitted by

DMT Batch-9

Under The Guidance of

Dr. M. H. M. Krishna Prasad

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Jawaharlal Nehru Technological University

Kakinada District, Andhra Pradesh – 533003

2022-2023

Jawaharlal Nehru Technological University

Kakinada District, Andhra Pradesh – 533003

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the seminar titled "COMPARISON OF APRIORI AND FP GROWTH" submitted by DMT Batch-9 **of Semester V is a bonafide account of the work done by him/her under our supervision, during the academic year 2022-2023**

SEMINAR GUIDE HEAD OF THE DEPARTMENT

ABSTRACT

- Association rule mining is one the most important concept that we use in our daily life. The two main algorithms that we deal in Association rule mining are Apriori and FP Growth Algorithms. So Through this paper let us understand the comparison between these two algorithms.
- Firstly, these algorithms are clearly explained each with an example and then there is a comparison section that actually gives you a good insight on the drawbacks of Apriori algorithm that were rectified by FP growth.
- At the end of this paper you will understand the benefits of using FP growth algorithm rather than Apriori Algorithm for a large dataset.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

The association rule learning is one of the very important concepts of Machine learning, and it is employed in Market Basket analysis, Web usage mining, continuous production**,** etc**.** Here market basket analysis is a technique used by the various big retailer to discover the associations between items. We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.

This paper deals with two of these association rule learning algorithms:

- Apriori algorithm
- FP growth algorithm

Now let's first know about Association rule learning and then dive into each algorithm in detail.

CHAPTER II

ASSOCIATION RULE MINING

- In a given set of transactions, finding interesting associations and relationships among large sets of data items is Association Rule Mining.
- Market Basket Analysis is one of the key techniques used by large relations to show association between items.
- Allows retailers to identify relationships between the items that people buy together.

TID	ITEMS
1	Bread , Milk
2	Bread ,Cake,Beer , Eggs
3	Milk ,Cake ,Beer , Coke
4	Bread ,Milk ,Cake , Beer
5	Bread ,Milk ,Cake , Coke

- On observing the given table, people who buy Cake also buy beer in most of the cases and people who buy milk may also buy coke

If you look at a lot of such transactions, we can observe certain patterns in the form of rules. We call them association rules.

- There are two parts in an association rule: the left hand side is an item set ; that means, a collection of items such as milk, bread etc and so as the right hand side.

- Examples of Association Rules:

{Cake} -> {Beer}

{Milk, Bread} -> {Eggs, Coke}

{Beer, Bread} -> {Milk}

- Can we find some association between these two items? In other words, how can we assume that whenever customer buys bread and milk, he most probably buys beer?

- So as to find out the association between the items we have to define three terms:

-
SUPPORT:

Fraction of transactions that contain the item sets X and Y .

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

- CONFIDENCE:

Measures how often items in Y appear in transactions that contain X.

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

- LIFT:
  Confidence of the rule divided by expected confidence

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$

- Lift near 1 indicates X and Y almost appear together as expected

- Lift >1 indicates X and Y appear together more than expected
- Lift<1 indicates X and Y appear together less than expected.

- NOTE:

For measuring associations for any transaction, we assume a minimum support and confidence . If it is not met, then the item set is discarded.

TID	ITEMS
1	Bread , Milk
2	Bread , Cake ,Beer , Eggs
3	Milk ,Cake ,Beer , Coke
4	Bread ,Milk ,Cake , Beer
5	Bread ,Milk ,Cake , Coke

- From the table, consider the association rule

{Milk, Cake}->{Beer}

Support= σ(Milk, Cake, Beer)/n = 2/5 =0.4

Confidence = σ(Milk, Cake, Beer)/σ(Milk, Cake)

= 2/3 = 0.67

Lift = supp(Milk, Cake, Beer)/supp(Milk, Cake)*supp(beer)

= 0.4/(0.6*0.6) = 1.11

APPLICATIONS:-

- This kind of pattern has huge commercial significance.
- Discounting of one product may increase the sales of other.
- Keeping the associated things together in a supermarket can increase sales.

- This just don’t increase the profit to the market but the customer can also buy things which are essential
- This type of association is not just present in market basket transactions but also in purchase of railway tickets, stocks etc.

CHAPTER III

APRIORI ALGORITHM

- This algorithm is also called as FREQUENT PATTERN MINING.

DEFINITION OF APRIORI ALGORITHM

- APRIORI - the basic algorithm for finding frequent itemsets.
- The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. Apriori is designed to operate on database containing transactions.
- Example: Collections of items bought by customers.
- Apriori employs an iterative approach known as a level-wise search, where k item sets are used to explore (k + 1) item sets.

WHAT ACTUALLY APRIORI SAYS?

The probability that item I is not frequent is if :

- P(I) < minimum support threshold, then I is not frequent.
- P (I+A) < minimum support threshold, then I+A is not frequent, where A also belongs to itemset.
- It follows Antimonotone property.
- The steps followed in apriori algorithm of data mining are:

➢ Join step

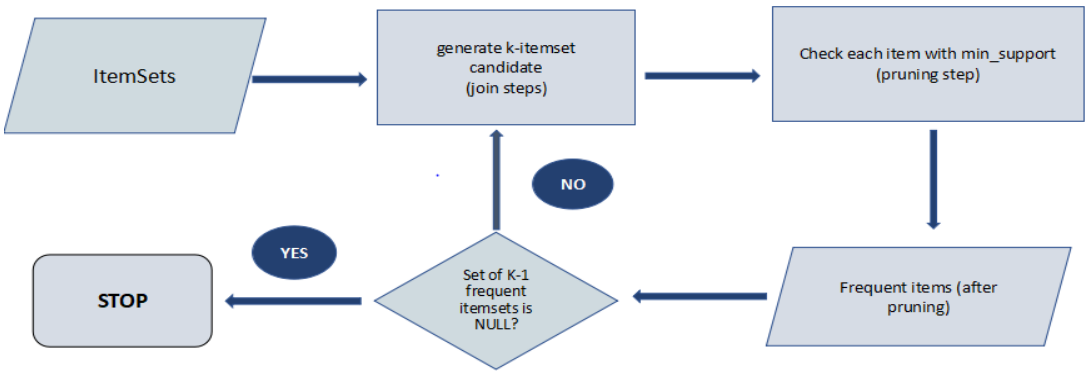
➢ Prune step

KEY CONCEPTS

- Frequent Itemsets : All the sets which contain the item with the minimum support
- Apriori Property : Any subset of frequent itemset must be frequent.

- Anti - monotonicity : It means if a set cannot pass a test, all of its supersets will fail the same test as well.
- Join Step : This step generates (K+1) item set from K – itemsets by joining each item with itself
- Prune Step : This step scans the count of each item in the database. If the candidate item does not meet minimum support, then it is regarded as infrequent and thus it is removed. This step is performed to reduce the size of the candidate itemsets.

STEPS TO PERFORM APRIORI ALGORITHM



PROS AND CONS OF APRIORI ALGORITHM

PROS:

➢ This is the most simple and easy-to-understand and implement.

➢ Join and Prune steps are easy to implement on item sets in database.

CONS:

➢ It scans the database multiple times for generating candidate sets.

➢ It requires large memory space due to large number of candidate generation.

CHAPTER IV

FP GROWTH ALGORITHM

- The FP-Growth Algorithm is an alternative way to find frequent item sets without using candidate generations, thus improving performance.
- Divide and Conquer Strategy
- The core of this method is the usage of a special data structure named frequent-pattern tree (FP-tree), which retains the item set association information.
- This algorithm works as follows:
- First, it compresses the input database creating an FP-tree instance to represent frequent items.
- After this first step, it divides the compressed database into a set of conditional databases, each associated with one frequent pattern.
- Finally, each such database is mined separately**.**

FP TREE

- Compact data structure that stores quantitative information about frequent patterns in a database.
- Each transaction is read and then mapped onto a path in the FP-tree.
- As different transactions can have several items in common, their paths may overlap.
- In the FP tree, the root node represents null, while the lower nodes represent the item sets. The associations of the nodes with the lower nodes, that is, the item sets with the other item sets, are maintained while forming the tree.

STEPS TO BE FOLLOWED:

➢ Scan the Database to find the occurrences of the itemsets

➢ Construct the FP tree (root node -> NULL)

➢ Examine the first transaction. The itemset with the max count is taken at the top and so on. The branch of the tree is constructed with itemsets in descending order of count

➢ Similarly, the next transaction is examined. The itemsets are ordered in descending order of count. If any itemset of the transaction is already present in another branch, then this transaction branch would share a common prefix to the root.

➢ Also, the count of the itemset is incremented as it occurs in the transactions. The common node and new node count are increased by 1 as they are created and linked according to transactions.

➢ Mine the created FP tree. the lowest node is examined first, along with the links of the lowest nodes. The lowest node represents the frequency pattern length 1. From this, traverse the path in the FP Tree. This path or paths is called a conditional pattern base.
A conditional pattern base is a sub-database consisting of prefix paths in the FP tree occurring with the lowest node (suffix).

➢ Construct a Conditional FP Tree, formed by a count of itemsets in the path. The itemsets meeting the threshold support are considered in the Conditional FP Tree.

➢ Frequent Patterns are generated from the Conditional FP Tree.

◆ **EXAMPLE:**

The given data is a hypothetical dataset of transactions with each letter representing an item

TRANSACTION ID	ITEMS
T1	{E,K,M,N,O,Y}
T2	{D,E,K,N,O,Y}
T3	{A,E,K,M}
T4	{C,K,M,U,Y}
T5	{C,E,I,K,O,O}

◆ The frequency of each individual item is computed.

◆ Let the minimum support be 3.

◆ A Frequent Pattern set is built which will contain all the elements whose frequency is greater than or equal to the minimum support. These elements are stored in descending order of their respective frequencies. After insertion of the relevant items, the set L looks like this:-

ITEM	FREQUENCY
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	2
O	3
U	1
Y	3

L= {K : 5,

E : 4,

M : 3,

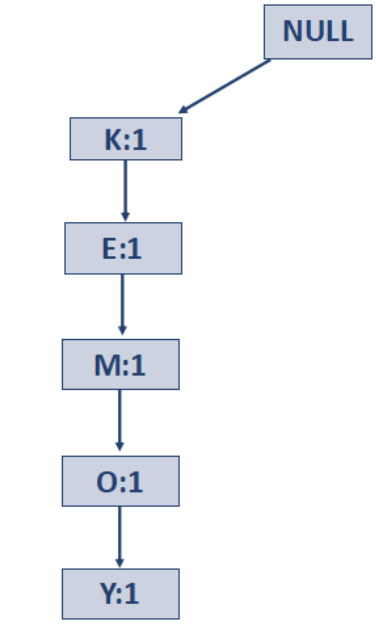
O : 3,

Y : 3}

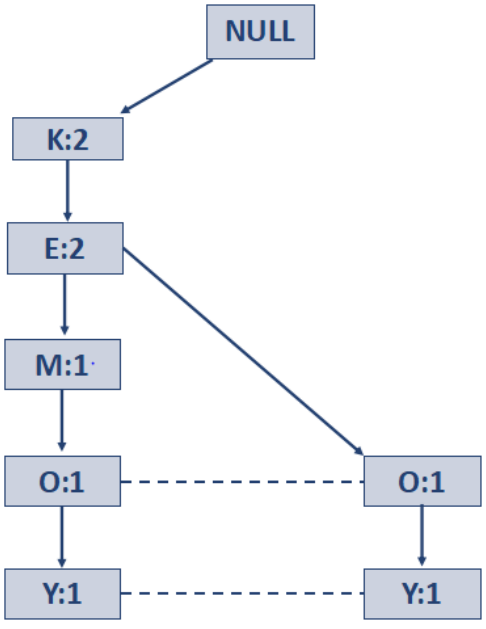
For each transaction, the respective Ordered-Item set is built.

TRANSACTION ID	ITEMS	ORDERED-ITEMSET
T1	{E,K,M,N,O,Y}	{K,E,M,O,Y}
T2	{D,E,K,N,O,Y}	{K,E,O,Y}
T3	{A,E,K,M}	{K,E,M}
T4	{C,K,M,U,Y}	{K,M,Y}
T5	{C,E,I,K,O,O}	{K,E,O}

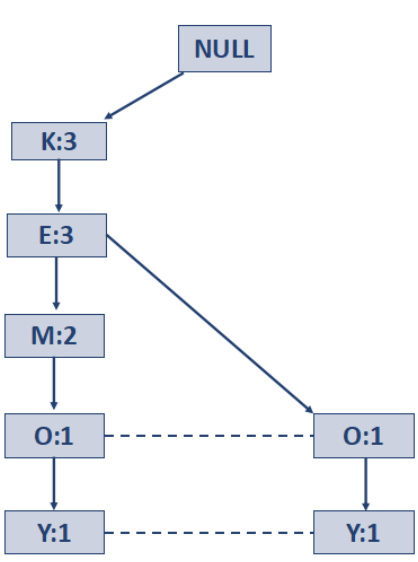
- Now, all the Ordered-Item sets are inserted into a Tree Data Structure
- Now, let's see how the FP tree is constructed using the above ordered set.



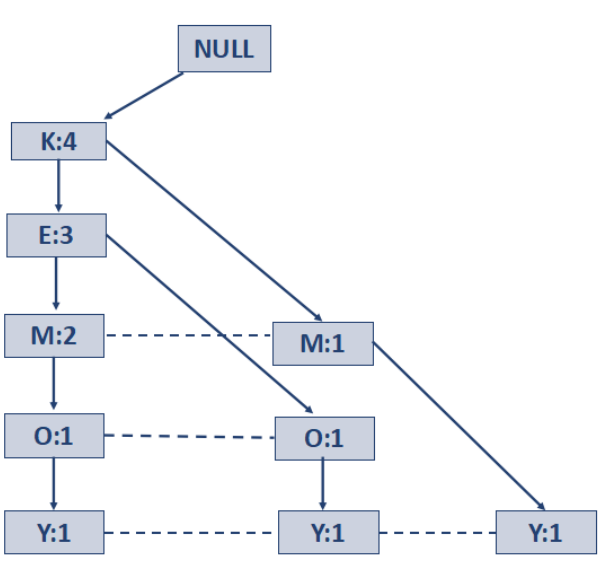
a) Inserting the set {K,E,M,O,Y}



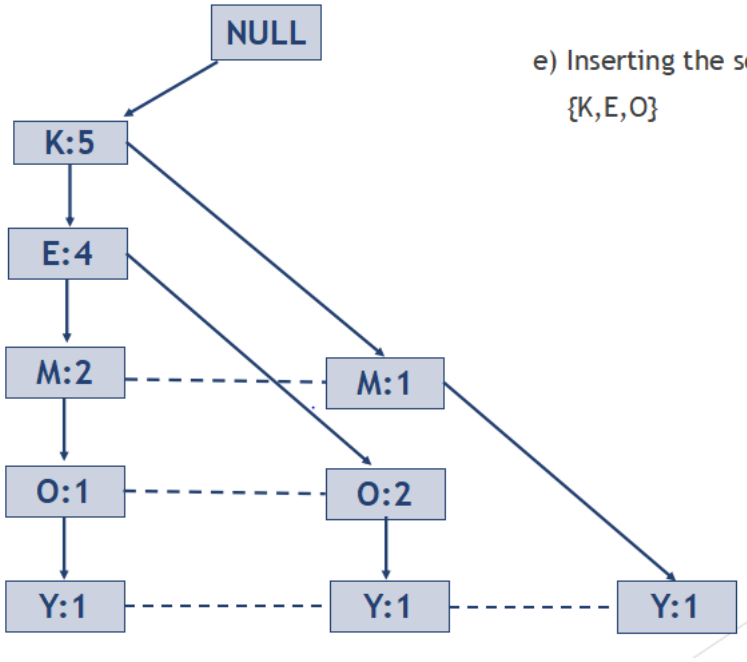
b) Inserting the set {K,E,O,Y}



c) Inserting the set {K,E,M}



d) Inserting the set {K,M,Y}



e) Inserting the set {K,E,O}

Now, for each item, the Conditional Pattern Base is computed.

ITEMS	CONDITIONAL PATTERN BASE
Y	{{K,E,M,O:1},{K,E,O:1},{K,M:1}}
O	{{K,E,M:1},{K,E:2}}
M	{{K,E:2},{K:1}}
E	{K:4}
K	

◆ Now, for each item, the Conditional Frequent Pattern Tree is built

ITEMS	CONDITIONAL PATTERN BASE	CONDITIONAL FREQUENT PATTERN TREE
Y	{{K,E,M,O:1},{K,E,O:1},{K,M:1}}	{K:3}
O	{{K,E,M:1},{K,E:2}}	{K,E:3}
M	{{K,E:2},{K:1}}	{K:3}
E	{K:4}	{K:4}
K		

◆ From the Conditional Frequent Pattern tree, the Frequent Pattern rules are generated by pairing the items of the Conditional Frequent Pattern Tree set to the corresponding to the item as given in the below table.

ITEMS	FREQUENT PATTERN GENERATED
Y	{<K,Y:3>}
O	{<K,O:3>,<E,O:3>,<E,K,O:3>}
M	{<K,M:3>}
E	{<E,K:4>}
K	

◆ For each row, two types of association rules can be inferred for example for the first row which contains the element, the rules K -> Y and Y -> K can be inferred. To determine the valid rule, the confidence of both the rules is calculated and the one with confidence greater than or equal to the minimum confidence value is retained.

TRANSACTION ID	ITEMS
T1	{E,K,M,N,O,Y}
T2	{D,E,K,N,O,Y}
T3	{A,E,K,M}
T4	{C,K,M,U,Y}
T5	{C,E,I,K,O,O}

◆ Let us consider the minimum confidence as 0.65

Some of the association rules that can be formed are

{K, O}->{E} (C=3/3=1)

{E}->{O} (C=3/4=0.75)

{M}->{K} (C=3/3=1)

ITEMS	FREQUENT PATTERN GENERATED
Y	{<K,Y:3>}
O	{<K,O:3>,<E,O:3>,<E,K,O:3>}
M	{<K,M:3>}
E	{<E,K:4>}
K	

CHAPTER V

DIFFERENCE BETWEEN APRIORI AND FP GROWTH

APRIORI	FP GROWTH
◆ Array based algorithm.	◆ Tree based algorithm.
◆ Uses Join and Prune technique.	◆ Constructs conditional frequent pattern tree and conditional pattern base from database which satisfy minimum support
◆ Uses a breadth-first search	◆ Uses a depth-first search
◆ Utilizes a level-wise approach where it generates patterns containing 1 item, then 2 items, then 3 items, and so on.	◆ Utilizes a pattern-growth approach means that, it only considers patterns actually existing in the database.
◆ Candidate generation is extremely slow. Runtime increases exponentially depending on the number of different items.	◆ Runtime increases linearly, depending on the number of transactions and items

◆ Scans the database multiple times for generating candidate sets.	◆ Scans the database only twice for constructing frequent pattern tree.
◆ Requires large memory space due to large number of candidate generation.	◆ Requires less memory space due to compact structure and no candidate generation.

CHAPTER VI

CONCLUSIONS

◆ Association rule mining plays an important role in our daily life. They are used in Supermarket arrangement, Recomender Systems, Web usage mining, etc. So one must be aware on application of the algorithm for mining these rules.

◆ After looking into all the differences between Apriori and Fp growth it can be easily understood that for a small dataset both algorithms have similar performance but as the size of the dataset increses FP growth algorithm performs way better than Apriori in terms of both space and time efficiency.

CHAPTER VII

BIBLIOGRAPHY

- **Data MINING concepts and techniques** By Jiawei Han, Micheline Kamber, Jain Pei.

- https://www.researchgate.net/publication/228913454_An_Implementation_of_the_FP-growth_Algorithm .

- https://analyticsindiamag.com/apriori-vs-fp-growth-in-market-basket-analysis-a-comparative-guide/