Chapter-III

1. **Data Mining :**
   - Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets.
   - Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data.
   - It is a multi-disciplinary skill that uses machine learning, statistics, AI and database technology.
   - The insights derived via Data Mining can be used for marketing, fraud detection, and scientific discovery, etc.
   - Data mining is also called as Knowledge discovery, Knowledge extraction, data/pattern analysis, information harvesting, etc.

2. **Types of Data**

Data mining can be performed on following types of data

- Relational databases
- Data warehouses
- Advanced DB and information repositories
- Object-oriented and object-relational databases
- Transactional and Spatial databases
- Heterogeneous and legacy databases
- Multimedia and streaming database
- Text databases
- Text mining and Web mining

3. **Data Mining Implementation Process**

- **Business understanding:**

In this phase, business and data-mining goals are established.

1. First, you need to understand business and client objectives.
2. You need to define what your client wants
3. Take stock of the current data mining scenario-resources, assumption, constraints.
4. A good data mining plan is very detailed

- **Data understanding:**

In this phase, sanity check on data is performed to check whether its appropriate for the data mining goals.

1. First, data is collected from multiple data sources available in the organization.
2. These data sources may include multiple databases, flat filer or data cubes. For example, table A contains an entity named cust_no whereas another table B contains an entity named cust-id.

- **Data preparation:**

In this phase, data is made production ready.

1. The data preparation process consumes about 90% of the time of the project.
2. The data from different sources should be selected, cleaned, transformed, formatted.
3. Following transformation can be applied

Data transformation:

Data transformation operations

i. **Smoothing:** It helps to remove noise from the data.
ii. **Aggregation:** Summary or aggregation operations are applied to the data. I.e., the weekly sales data is aggregated to calculate the monthly and yearly total.
iii. **Generalization:** In this step, Low-level data is replaced by higher-level concepts with the help of concept hierarchies. For example, the city is replaced by the county.
iv. **Normalization:** Normalization performed when the attribute data are scaled up or scaled down.
v. **Attribute construction**: these attributes are constructed and included the given set of attributes helpful for data mining.

The result of this process is a **final data set** that can be used in modeling.

- **Modelling**

In this phase, mathematical models are used to determine data patterns.

1. Based on the business objectives, suitable modeling techniques should be selected.
2. Create a scenario to test check the quality and validity of the model.
3. Run the model on the prepared dataset.
4. Results should be assessed by all stakeholders to make sure that model can meet data mining objectives.

- **Evaluation:**

In this phase, patterns identified are evaluated against the business objectives.

1. Results generated by the data mining model should be evaluated against the business objectives.
2. Gaining business understanding is an iterative process.
3. A go or no-go decision is taken to move the model in the deployment phase.

- **Deployment:**

In the deployment phase, you ship your data mining discoveries to everyday business operations.

1. The knowledge or information discovered during data mining process should be made easy to understand for non-technical stakeholders.
2. A detailed deployment plan, for shipping, maintenance, and monitoring of data mining discoveries is created.
3. A final project report is created with lessons learned and key experiences during the project. This helps to improve the organization's business policy.

4. **Data Mining Techniques**

1. Classification:

This data mining method helps to classify data in different classes.

2. Clustering:

Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

3. Regression:

Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

4. Association Rules:

This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

5. Outer detection:

This type of data mining technique refers to observation of data items in the dataset which do not match an expected pattern or expected behaviour. This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier mining.

6. Sequential Patterns:

This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period.

7. Prediction:

Prediction has used a combination of the other data mining techniques like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in a right sequence for predicting a future event.

5. **Challenges of Implementation of Data mine:**

i. Skilled Experts are needed to formulate the data mining queries.
ii. Overfitting: Due to small size training database, a model may not fit future states.
iii. Data mining needs large databases which sometimes are difficult to manage
iv. Business practices may need to be modified to determine to use the information uncovered.
v. If the data set is not diverse, data mining results may not be accurate.

6. **Data mining Examples:**

Example 1:

Consider a marketing head of telecom service provides who wants to increase revenues of long distance services.

Example 2:

A bank wants to search new ways to increase revenues from its credit card operations. They want to check whether usage would double if fees were halved.

7. **Data Mining Tools**

Following are 2 popular Data Mining Tools widely used in Industry

i. R-language
ii. Oracle Data Mining:
iii. Rapid Miner (erstwhile YALE):
iv. WEKA

Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.

8. **Benefits of Data Mining:**

- Get knowledge-based information.
- In order to make the profitable adjustments in operation and production.
- Cost-effective and efficient solution compared to other statistical data applications.
- Data mining helps with the decision-making process.
- Facilitates automated prediction of trends ,behaviors as well as automated discovery of hidden patterns.
- It can be implemented in new systems as well as existing platforms
- It is the speedy process which makes it easy for the users to analyze huge amount of data in less time.

9. **Disadvantages of Data Mining**

- There are chances of companies may sell useful information of their customers to other companies for money. For example, American Express has sold credit card purchases of their customers to the other companies.
- Many data mining analytics software is difficult to operate and requires advance training to work on.
- Different data mining tools work in different manners due to different algorithms employed in their design. Therefore, the selection of correct data mining tool is a very difficult task.
- The data mining techniques are not accurate, and so it can cause serious consequences in certain conditions.

10. **Where is Data Mining used?**

Applications	Usage
Communications	Data mining techniques are used in communication sector to predict customer behavior to offer highly targetted and relevant campaigns.
Insurance	Data mining helps insurance companies to price their products profitable and promote new offers to their new or existing customers.

Education	Data mining benefits educators to access student data, predict achievement levels and find students or groups of students which need extra attention. For example, students who are weak in maths subject.
Manufacturing	With the help of Data Mining Manufacturers can predict wear and tear of production assets. They can anticipate maintenance which helps them reduce them to minimize downtime.
Banking	Data mining helps finance sector to get a view of market risks and manage regulatory compliance. It helps banks to identify probable defaulters to decide whether to issue credit cards, loans, etc.
Retail	Data Mining techniques help retail malls and grocery stores identify and arrange most sellable items in the most attentive positions. It helps store owners to comes up with the offer which encourages customers to increase their spending.
Service Providers	Service providers like mobile phone and utility industries use Data Mining to predict the reasons when a customer leaves their company. They analyze billing details, customer service interactions, complaints made to the company to assign each customer a probability score and offers incentives.
E-Commerce	E-commerce websites use Data Mining to offer cross-sells and up-sells through their websites. One of the most famous names is Amazon, who use Data mining techniques to get more customers into their eCommerce store.
Super Markets	Data Mining allows supermarket's develope rules to predict if their shoppers were likely to be expecting. By evaluating their buying pattern, they could find woman customers who are most likely pregnant. They can start targeting products like baby powder, baby shop, diapers and so on.
Crime Investigation	Data Mining helps crime investigation agencies to deploy police workforce (where is a crime most likely to happen and when?), who to search at a border crossing etc.
Bioinformatics	Data Mining helps to mine biological data from massive datasets gathered in biology and medicine.

Association Rule Mining:

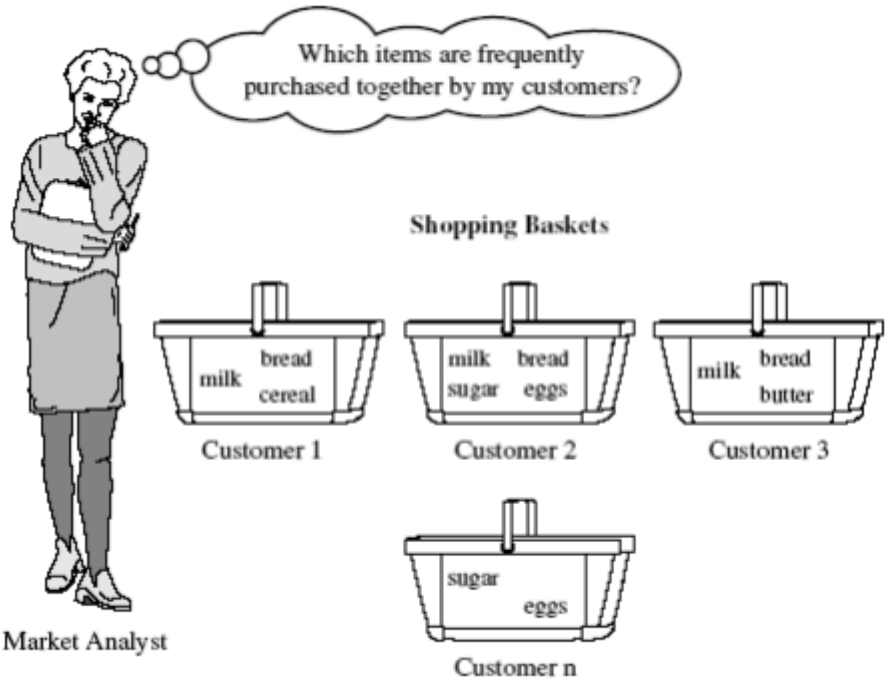
Wal-Mart, the world's largest retailer- popped a most unexpected correlation:

sales of diapers and beer.

1. On Fridays the men figured they deserved a six-pack of beer for their trouble.

2. Because there's no time left to go to a bar, take beer home with them.

- Association rule mining is a popular and well researched method for discovering interesting relations between variables in large databases.
- It is introduced in 1993 by agrawal.
- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction



The problem of mining association rules can be stated as follows: Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of **items**. Let $T = (t_1, t_2, \ldots, t_n)$ be a set of **transactions** (the database), where each transaction t_i is a set of items such that $t_i \subseteq I$. An **association rule** is an implication of the form, $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$

X (or Y) is a set of items, called an **itemset**.

A transaction $t_i \in T$ is said to **contain** an itemset X if X is a subset of t_i (we also say that the itemset X **covers** t_i). The **support count** of X in T(denoted by $X.count$) is the number of

Transactions in T that Contains X .The strength of a rule is measured by its support and confidence.

Support: The support of a rule, $X \rightarrow Y$, is the percentage of transactions in T that contains $X \cup Y$, and can be seen as an estimate of the probability, $\Pr(X \cup Y)$. The rule support thus determines how frequent the rule is applicable in the transaction set T. Let n be the number of transactions in T. The support of the rule $X \rightarrow Y$ is computed as follows:

support $= (X \cup Y).count / n$

Support is a useful measure because if it is too low, the rule may just occur due to chance. Furthermore, in a business environment, a rule **covering** too few cases (or transactions) may not be useful because it does not make business sense to act on such a rule (not profitable).

Confidence: The confidence of a rule, $X \rightarrow Y$, is the percentage of transactions in T that contain X also contain Y. It can be seen as an estimate of the conditional probability, $\Pr(Y \mid X)$. It is computed as follows:

confidence $= (X \cup Y).count / X.count$

Confidence thus determines the **predictability** of the rule. If the confidence of a rule is too low, one cannot reliably infer or predict Y from X. A rule with low predictability is of limited use.

- **Itemset**
  - A collection of one or more items
    - Example: {Milk, Bread, Diaper}
  - k-itemset
    - An itemset that contains k items
- **Support count (σ)**
  - Frequency of occurrence of an itemset
  - E.g. σ({Milk, Bread,Diaper}) = 2
- **Support**
  - Fraction of transactions that contain an itemset
  - E.g. s({Milk, Bread, Diaper}) = 2/5
- **Frequent Itemset**
  - An itemset whose support is greater than or equal to a *minsup* threshold

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- **Association Rule**
  - An implication expression of the form $X \rightarrow Y$, where X and Y are itemsets
  - Example: {Milk, Diaper} → {Beer}

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- **Rule Evaluation Metrics**
  - Support (s)
    - Fraction of transactions that contain both X and Y
  - Confidence (c)
    - Measures how often items in Y appear in transactions that contain X

Example:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Rules:

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

Observations:

- All the above rules are binary partitions of the same itemset: {Milk, Diaper, Beer}
- Rules originating from the same itemset have identical support but can have different confidence

Example2: Example database with 4 items and 5 transactions

Transaction ID	milk	bread	butter	beer
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

{Butter, Bread}⟹ {Milk}

S= σ (Butter, Bread,Milk)/|T|=1/5=0.2

C=Conf(X⟹Y)=Sup(X∪Y)/Sup(X)

=σ(Butter, Bread,Milk)/σ (Butter, Bread)

=0.2/0.2=1

Lift: The lift of rule is defind as

Lift(X⟹Y)= Sup(X∪Y) / Sup(X). Sup(Y).

In above example

Lift({Butter, Bread} ⟹ {Milk})= Sup({Butter, Bread}∪{Milk}) / Sup{Butter, Bread}* Sup {Milk}

=0.2/0.2*0.4

=2.25

Lift({Milk, Bread} ⟹ {Butter})=Sup({Milk, Bread }∪{ Butter }) /Sup{ Milk, Bread }* Sup { Butter }

=0.2/0.4*0.4

=1.25

- **Apriori Algorithm**

1. Apriori algorithm is a classical algorithm in data mining. It is used for mining frequent itemsets and relevant association rules.
2. It is devised to operate on a database containing a lot of transactions, for instance, items brought by customers in a store.
3. This algorithm, introduced by R Agrawal and R Srikant in 1994 has great significance in data mining.
4. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties.
5. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets.

Apriori Property –"All non-empty subset of frequent itemset must be frequent."

Apriori algorithm – The Theory

Three significant components comprise the apriori algorithm. They are as follows.

- Support
- Confidence
- Lift

Consider the following dataset and we will find frequent item sets and generate association rules for them.

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

minimum support count is 0.2 ie (0.2*No Transactions=0.2*9=1.8 ie 2 Minimum Transactions)
minimum confidence is 60%

Step-1: K=1
(I) Create a table containing support count of each item present in dataset – Called **C1(candidate set)set**

Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

(II) compare candidate set item's support count with minimum support count(here min_support=2 if support_count of candidate set items is less than min_support then remove those items). This gives us itemset L1.

Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

Step-2: K=2

- Generate candidate set C2 using L1 (this is called join step). Condition of joining $L_{k-1}$and L_{k-1} is that it should have (K-2) elements in common.
- Check all subsets of an itemset are frequent or not and if not frequent remove that itemset.(Example subset of{I1, I2} are {I1}, {I2} they are frequent.Check for each itemset)
- Now find support count of these itemsets by searching in dataset.

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I4	1
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I3,I4	0
I3,I5	1
I4,I5	0

(II) compare candidate (C2) support count with minimum support count(here min_support=2 if support_count of candidate set item is less than min_support then remove those items) this gives us itemset L2.

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I2,I5	2

Step-3:

- Generate candidate set C3 using L2 (join step). Condition of joining L_{k-1} and L_{k-1} is that it should have (K-2) elements in common. So here, for L2, first element should match.
  So itemset generated by joining L2 is {I1, I2, I3}{I1, I2, I5}{I1, I3, i5}{I2, I3, I4}{I2, I4, I5}{I2, I3, I5}

- Check if all subsets of these itemsets are frequent or not and if not, then remove that itemset.(Here subset of {I1, I2, I3} are {I1, I2},{I2, I3},{I1, I3} which are frequent. For {I2, I3, I4}, subset {I3, I4} is not frequent so remove it. Similarly check for every itemset)

- find support count of these remaining itemset by searching in dataset.

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

(II) Compare candidate (C3) support count with minimum support count(here min_support=2 if support_count of candidate set item is less than min_support then remove those items) this gives us itemset L3.

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

Step-4:

- Generate candidate set C4 using L3 (join step). Condition of joining L_{k-1} and L_{k-1}(K=4) is that, they should have (K-2) elements in common. So here, for L3, first 2 elements (items) should match.

- Check all subsets of these itemsets are frequent or not (Here itemset formed by joining L3 is {I1, I2, I3, I5} so its subset contains {I1, I3, I5}, which is not frequent). So no itemset in C4

- We stop here because no frequent itemsets are found further

- Thus, we have discovered all the frequent item-sets. Now generation of strong association rule comes into picture. For that we need to calculate confidence of each rule.
- **Confidence –**
  A confidence of 60% means that 60% of the customers, who purchased milk and bread also bought butter.
  - Confidence(A->B)=Support_count(A∪B)/Support_count(A)
- So here, by taking an example of any frequent itemset, we will show the rule generation.
  Itemset {I1, I2, I3} //from L3
  SO rules can be
  [I1^I2]=>[I3] //confidence = sup(I1^I2^I3)/sup(I1^I2) = 2/4*100=50%
  [I1^I3]=>[I2] //confidence = sup(I1^I2^I3)/sup(I1^I3) = 2/4*100=50%
  [I2^I3]=>[I1] //confidence = sup(I1^I2^I3)/sup(I2^I3) = 2/4*100=50%

[I1]=>[I2^I3] //confidence = sup(I1^I2^I3)/sup(I1) = 2/6*100=33%
[I2]=>[I1^I3] //confidence = sup(I1^I2^I3)/sup(I2) = 2/7*100=28%
[I3]=>[I1^I2] //confidence = sup(I1^I2^I3)/sup(I3) = 2/6*100=33%

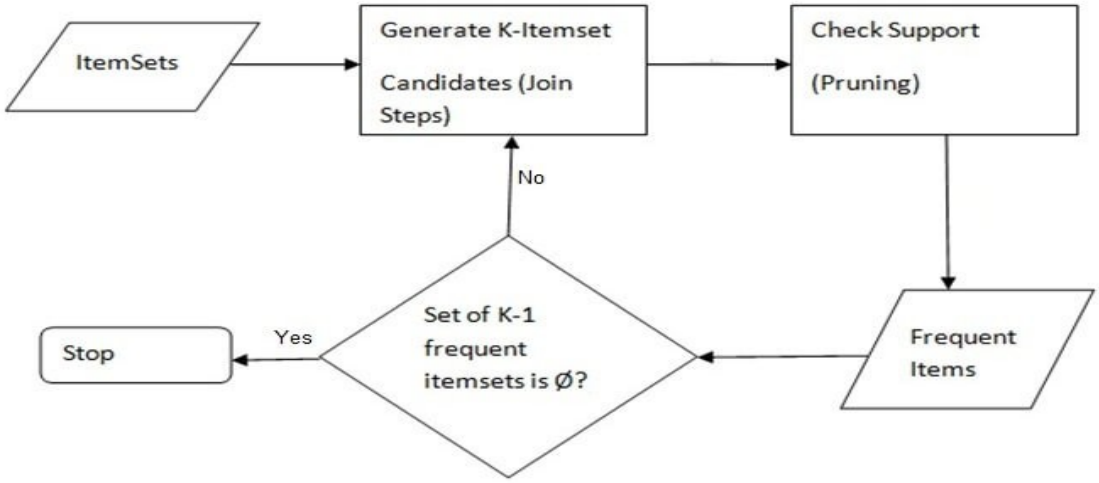
- So if minimum confidence is 50%, then first 3 rules can be considered as strong association rules.

➢ The entire algorithm can be divided into two steps:

Step 1: Apply minimum support to find all the frequent sets with k items in a database.

Step 2: Use the self-join rule to find the frequent sets with k+1 items with the help of frequent k-itemsets. Repeat this process from k=1 to the point when we are unable to apply the self-join rule.

This approach of extending a frequent itemset one at a time is called the "bottom up" approach.



- Join Step: C_k is generated by joining L_{k-1} with itself
- Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
- Pseudo-code :C_k: Candidate itemset of size k
  L_k: frequent itemset of size k

```
L1 = {frequent items};
for (k = 1; Lk !=∅; k++) do begin
   Ck+1 = candidates generated from Lk;
   for each transaction t in database do
         increment the count of all candidates in Ck+1
          that are contained in t
   Lk+1 = candidates in Ck+1 with min_support
   end
return ∪k Lk;
```

Apriori Algorithm – Pros

- Easy to understand and implement
- Can use on large itemsets

Apriori Algorithm – Cons

- At times, you need a large number of candidate rules. It can become computationally expensive.
- It is also an expensive method to calculate support because the calculation has to go through the entire database.

Apriori Algorithm – Limitations

- The process can sometimes be very tedious.

How to Improve the Efficiency of the Apriori Algorithm?

Use the following methods to improve the efficiency of the apriori algorithm.

- **Transaction Reduction –** A transaction not containing any frequent k-itemset becomes useless in subsequent scans.
- **Hash-based Itemset Counting –** Exclude the k-itemset whose corresponding hashing bucket count is less than the threshold is an infrequent itemset.

There are other methods as well such as partitioning, sampling, and dynamic itemset counting.

Association Rule Mining Algorithm	Advantages	Disadvantages
AIS	1. An estimation is used in the algorithm to prune those candidate itemsets that have no hope to be large. 2. It is suitable for low cardinality sparse transaction database.	1. It is limited to only one item in the consequent. 2. Requires Multiple passes over the database. 3. Data structures required for maintaining large and candidate itemsets is not specified.
Apriori	1. This algorithm has least memory consumption. 2.Easy implementation. 3. It uses Apriori property for pruning therefore, itemsets left for further support checking remain less.	1. It requires many scans of database. 2. It allows only a single minimum support threshold. 3. It is favourable only for small database. 4. It explains only the presence or absence of an item in the database.
FP- growth	1. It is faster than other association rule mining algorithm. 2. It uses compressed representation of original database. 3. Repeated database scan is eliminated.	1. The memory consumption is more. 2. It cannot be used for interactive mining and incremental mining. 3. The resulting FP-Tree is not unique for the same logical database