Data Mining

Unit-1

Data Mining :-

Data Mining is defined as procedure of extracting information from huge sets of data.

(or)

It is also defined as mining knowledge from data.

Data :-

Data is defined as facts (or) figures (or) information that's stored in (or) used by computer.

Types of Data :-

There are three types of data. They are :

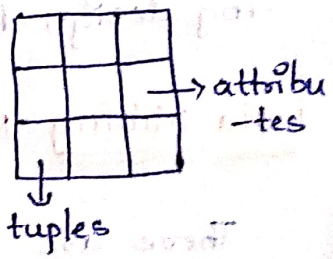
1. Database
2. Dataware House
3. Transactional database

1. Database Data (RDBMS) :

It is set of tables, which contains rows and columns.

Row Represents the tuples

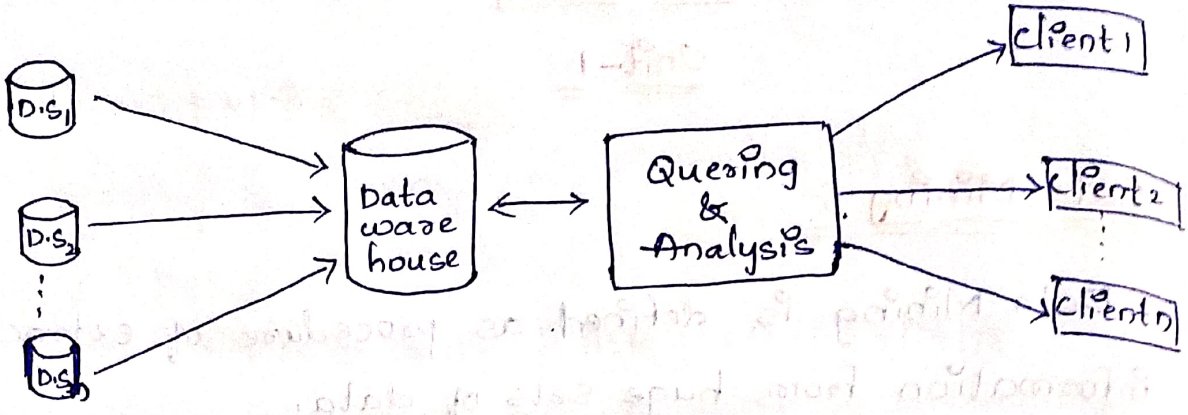
Column Represents the attributes.

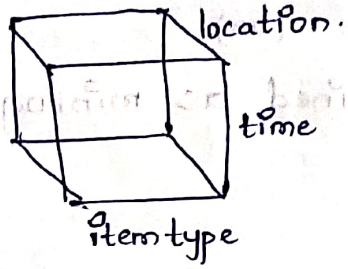


2. Dataware House :

Collection of data integrated from different sources with quering and decision making on data.

In data ware house, data is stored in dimension -al structure (data cube) where each dimension is each attribute.





3. Transactional database :

Each record is called as transaction (sales, flight booking, etc).

Transaction has transaction ID, List of other items making transactions.

From transaction DB, we can mine frequent patterns

Other types of data :

Sequence data, spatial data, data streams, engineering design data, web data, etc.

Data Mining Functionalities :-

There are 5 functionalities. They are :

1. Concept / class descriptions
2. Mining frequent patterns, association & correlation
3. Classification & Regression for predictive Analysis.
4. Cluster Analysis.
5. Outlier Analysis.

① Class description / concept:

Data is associated with class/concept.

Descriptions can be done in two ways.

a) Data Characterisation

b) Data Discrimination

a) Data Characterisation:

It refers to the summary of the class/concept.

o/p → General overview.

b) Data Discrimination:

It compares the common features of class/concept.

o/p → barcharts, curves, etc.

② Mining frequent patterns, associations & correlation:

Frequent Patterns:

Things which are found most commonly in data.

Frequent itemsets

Frequent Subsequence

Frequent Substructure.

Association Analysis:

It is a way of identifying the relation between various items.

Ex: used to determine sales of items that are frequently purchased together.

Correlation Analysis:

It is Mathematical technique.

It shows how strongly pair of attributes are related together.

Ex: Tall people tend to have more weight.

③ Classification and Regression for predictive Analysis

Classification :

Process of finding a model that distinguishes data items.

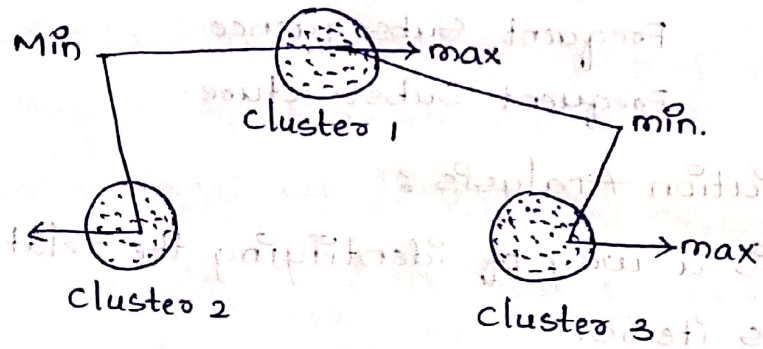
Decision tree is used for classification.

Regression :

It is statistical Methodology that is used for numeric prediction of missing data.

④ Cluster Analysis :

The data items are clustered based on the principles of maximising the intraclass similarity and minimising the interclass similarity.



⑤ Outlier Analysis :

Among the data items in a DB, there may be some items which do not follow the general behaviour of data.

Those data items are called as Outlier (noise/exceptions)

Intrestingness of patterns :-

In a data mining system, every day millions of data patterns are generated.

Among all these patterns generated, how many are really intresting ?

Actually a small fraction of patterns generated would be of intrest to any given user.

This raises three questions :

1. What makes patterns intresting ?

   A pattern is intresting if it is

   - Easily understood by humans
   - Valid on new/test data.
   - It is potentially useful.

2. Can data mining system generate all of the intresting patterns ?

   - It refers to completeness of data mining system.

In reality it is not possible for a data mining system to generate all intresting patterns.

3. Can data mining system generates only intresting patterns ?

   - It refers to optimization of a data mining system.
   - It generates only intresting patterns, because it becomes easy and efficient for the user
   - Time is saved.

Classification of data mining Systems :-

Classification :

Process of finding a model that distinguishes the data items.

Decision trees are used for classification.

- Data Mining is every where and Anywhere. It is used by many users.
- Data mining systems are classified based on several criteria.

1. Classification based on mined databases :

It is based on database that is been mined

- Relation
- Transactional
- object relational
- Datawarehouse

2. Classification based on type of knowledge mined :

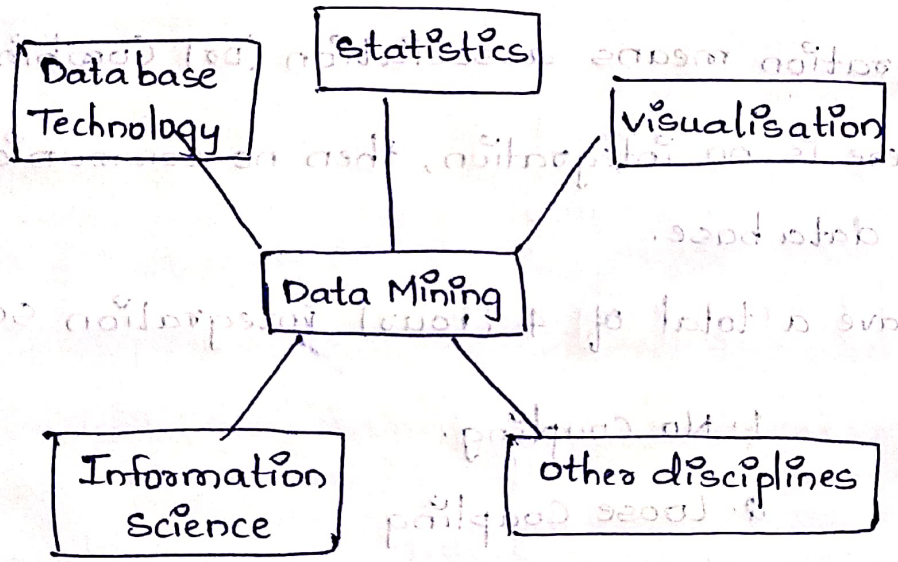
- characterization
- Discrimination
- Association and correlation Analysis
- Classification
- Prediction
- Outlier Analysis
- evolution Analysis.

3. Classification based on kinds of techniques used :

- ML
- Statistics

- Neural Networks
- Pattern recognition
- dataware house Oriented technique, etc.

4. Classification based on application adapted :
   - Finance
   - Telecommunication
   - DNA
   - Stock markets
   - emails, etc.



Data Mining Task Primitives :-

- A Data Mining Task is represented in the form of a datamining query.
- Data Mining query is defined as Data mining task primitives.
- It will allow user to interactively communicate with the data mining system.
- There are 5 data mining primitives. They are :

1. Set of task relevant data to be mined.
2. Specifies the kind of knowledge to be mined.
3. The background knowledge to be used in discovery process.
4. The intrestingness Measures and thresholds for pattern evaluation.
5. The expected representation for visualizing the discovered patterns.

Integration of Data mining System with a Data ware house :-

- Integration means association (or) Combining.
- If there is no integration, then no communication with data base.
- We have a total of 4 (four) integration schemas.

1. No Coupling.
2. Loose Coupling
3. Semitight Coupling
4. Tight Coupling.

1) No Coupling :

- Data mining system will not use any function.
- That is there is no communication with Database.
- Data mining system directly communicates with other storage methods to get data.

2) Loose coupling :

- Data Mining system will use some of the functionalities. (only upto some extent).
- It is better than No Coupling.
- It is suitable for small data sets.

3) Semi Tight Coupling :

- Data mining system is linked to the database but not Completely.
- Some of the data mining primitives are also implemented in database.

4) Tight Coupling :

- Data mining system is completely linked to the database.
- It is most efficient among all.
- The data base system is fully integrated ~~with~~ in such a way that it becomes a part of the data mining system.
- It is efficient and optimised implementation of data mining.

Major Issues in Data Mining :-

1. Mining different kinds of knowledge in data base.
2. Interactive mining of knowledge at multiple levels of abstractions.
3. Incorporation of background knowledge.
4. Presentation and visualization of data mining results.

5. Handling noisy / incomplete data.

6. Efficiency and scalability of data mining algorithm

Data Preprocessing :-

The process of transforming raw data into an understandable format.

- 4 major tasks.

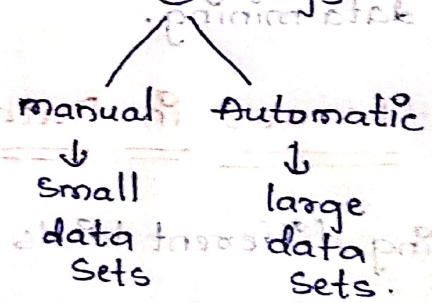
1. Data cleaning
2. Data Integration
3. Data Reduction
4. Data Transformation.

(1) Data cleaning :-

Process of removal of incorrect, incomplete, inaccurate data, also replaces missing data.

a) Handling missing values :

- In place of missing values, we can replace with "NA", with mean values, with median values.
- Sometimes replaced with most probable values
- missing values can be filled in (2) ways



b) Handling noisy data :

Noise data is inconsistent / error data.

Methods to handle :

(i) Binning :

First, data is sorted. Then sorted data is stored in bins.

③ methods to handle data in bins.

- smoothing by bin mean
- smoothing by bin median
- smoothing by bin boundary.

(ii) Regression:

Numerical prediction of data.

(iii) clustering:

similar data items are grouped at one place dissimilar items - outside the cluster.

② Data Integration:-

Multiple heterogeneous sources of data are combin-ed into single data set.

② types of data integration.

a) Tight coupling:

Data mining system is completely linked into the database.

b) Loose coupling:

Data mining system uses some of the functionali-ties (only to upto some extent).

③ Data Reduction:-

Volume of data is reduced to make analysis easier.

Methods:

a) Dimensionality Reduction:

Reduces no. of input variables in the dataset.

b) Data cube Aggregation:

Data is combined to construct a data cube.

c) Attribute subset selection:

Highly relevent attributes should be used.

d) Numerosity Reduction:

Here, we store only model of data instead of entire data.

④ Data Transformation:-

Data is transformed into appropriate form suitable for mining process.

methods:

a) Normalization:

Done in order to scale data values in specified range (-1.0 to 1.0 (or) from 0 to 1)

b) Attribute selection:

New attributes are selected by using old ones.

c) Discritization:

raw values are replaced by interval levels.

d) Concept Hierarchy Generation:

attributes are converted from low level to high level.

Ex: city → country.