2.11.23

Data Warehousing and Data Mining

- Data Mining- Mining knowledge from large amounts of data

- Mining of gold from rocks or sand

- Knowledge mining from DB, knowledge of extraction data pattern/ analysis.

Imp Qn

~~Knowledge discovery in databases~~

Knowledge discovery in databases (KDD)

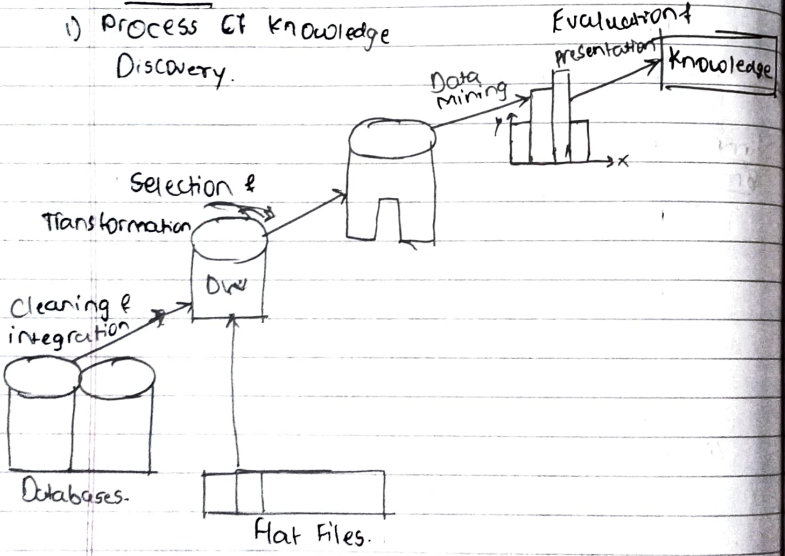
1) Data cleaning- To remove noise and inconsistent data
2) Data integration: Where multiple data sources may to combined
3) Data selection: Where data relevant to analysis are retrieved from databases.
4) Data transformation: Where data transformation is do. Consolidated in 2 forms for mining by performing summarizing or aggregation approach
5) Data Mining: Where inteligent methods are Applied for ~~det~~ extracting data patterns.

6) Pattern evaluation - to identify truely interesting patterns representing ~~patterns~~ knowledge

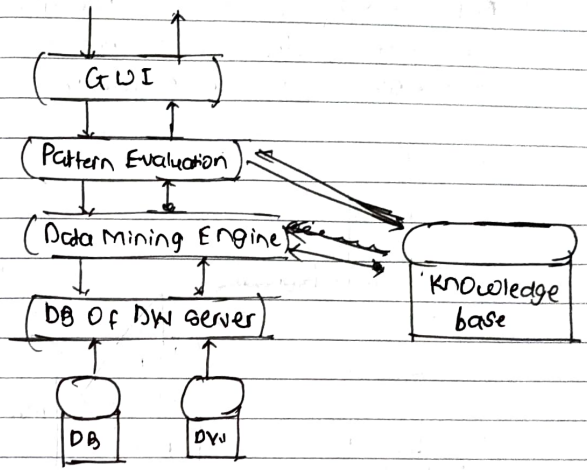
7) Knowledge presentation - where knowledge and visualization represent the mind knowledge to users

Diagram:

i) Process of knowledge Discovery.



Eg: Architecture of typical system.



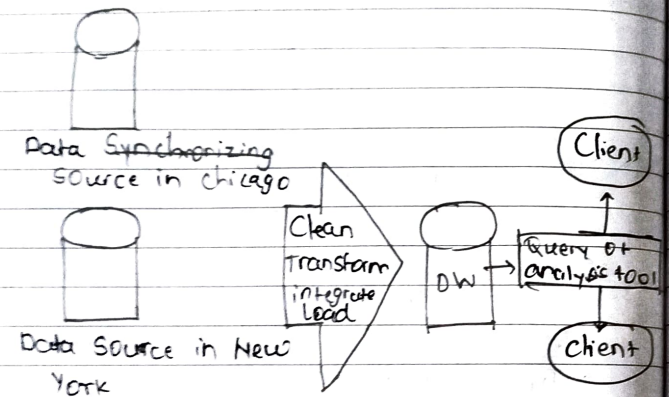
* Types of Data

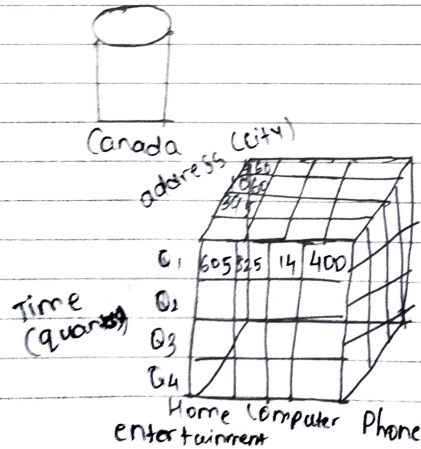
i) Relational Databases: Database Management systems consists of collection of iterelated data called as database. A relational database is a collection of tables each of which is assigned a unique name. Each table consist the set of attributes and stores large set of tuples.

Customer

Cust Id	Name	Address	Age	Income
101	AAA	Canada	21	$2700.

2) Data Warehouse: It is modeled by multi dimensional structure where each dimension corresponds to an attribute in the schema





Eg: Multidimensional data cube

3) Transactional databases: Transactional database consist of a file where each record represents transaction. A transaction includes transaction identity no which is unique, and list of items making up the transaction. it ~~enter~~ contains other info such as transaction Id, no branch, sales etc.

Sales

Trans-Id	List Of Items-Id
T100	T1, T3, T8
T200	T2, T4, T6
.	.
.	.
.	.

4) Advanced database Applications: Database Applications include

1) Handling special data
2) Engineering design data
3) Hypertext ~~Multimedia data~~
4) Time related data
5) WWW
6) Multimedia data

5) Object Oriented Databases:

It is based on OOPs where each entity is considered as objects. Objects can be individual employee, customer or item. Data and code are related to an object encapsulated in single unit.

- Spacial DB
- Temporal DB and Transeries DB
- Text DB and Multimedia DB
- Heterogenous DB
- WWW

12 marks

DM Functionalities - Kinds of patterns can be mined

- Data Mining functionalities
  1) Descriptive
  2) Predictive

1) Concept / class Description : Characterization
2) Association Analysis.
   a) It is the discovery of association rules showing attribute value condition that occur frequently together in a given set of data

a) $age(X, "20\ldots29") \wedge income(X, "20K\ldots29K") \Rightarrow buys(X, "CD\ player")$ [support = 2%, confidence = 60%]

b) $contains(T, "computer") \Rightarrow contains(T, "software")$ [support = 1%, confidence = 50%]

3) Classification and Prediction
   - Distinguish data classes / concepts.
   - Training data.
   - (IF-THEN) Rules

Eg: Good Response, Mild Response and Non-Response
   - All electronics company.

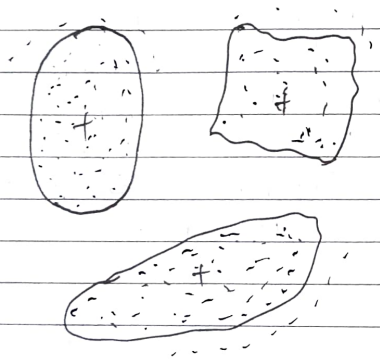
4) Cluster Analysis
   - Analysis data objects without consulting a known class label
   - Class labels are not present in training data

Eg.

	Question	Label
1	What is cluster analysis	Precriptive
2	Water content in bottom	Numeric
3	Where AIT is located	Location

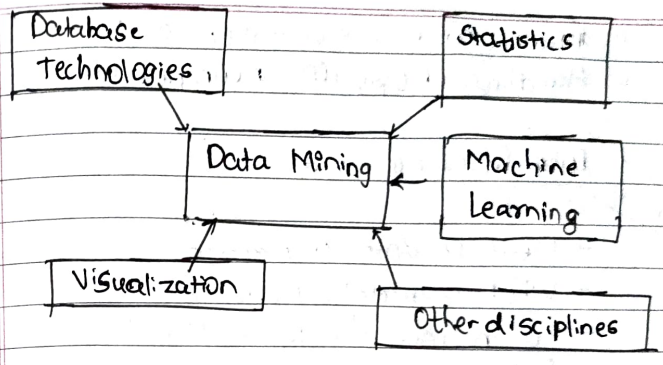
* unsupervised → without label
* supervised → with label

Defⁿ The Objects are clustered or grouped based on the principle of maximising intraclass similiarility and minimizing the interclass similiarility. The cluster of Objects are formed so that Objects within a cluster have high similiarity in comparison with one another and dissimiliar Obj. in other clusters.



⑤ Outliers Analysis.
- Don't comply with the general behaviour

* Classification of Data Mining System.



1) Classification according to databases Mined.

2) Classification according to Knowledge Mined
↳ Based on characterization, etc

3) Classification according to Techniques Mined
↳ ML, Data Warehouse, Visualization.

4) Classification according to applications adapted
↳ Finance, Stock Market, Weather Forecasting

* Major Issues in Data Mining
1) Mining different kinds of knowledge in DB.
2) Interactive mining of knowledge at multiple levels of abstraction.
3) Data Mining query languages.

4) Present and visualization of DM results
5) Handling noisy or incomplete data.

Data Processing

Lemarks

- Need of data preprocessing
- What is a data processing
- To improve the quality of data
- Data pre-processing techniques.

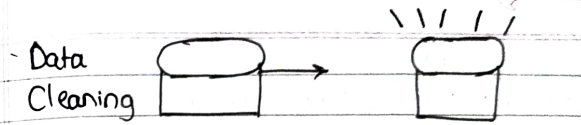
Data cleaning, Data Integration, Data Transformation, and Data Reduction.

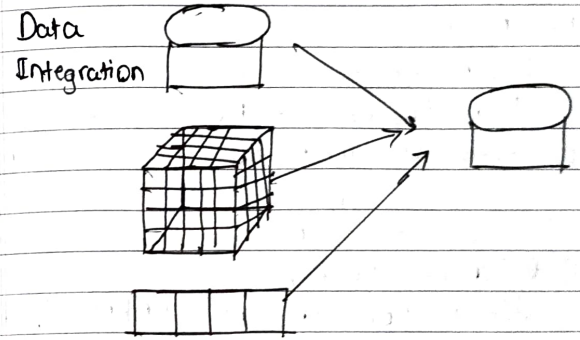
1) Data Cleaning is used to remove noice and correct inconsistent data

2) Data integration: Merges data ~~into a~~ from coherent data multiple sources into coherant data to store such as a Data warehouse.

3) Data Transformation: Normalization may be applied to improve efficiency & accuracy of mining algos.

4) Data Reduction: Can reduce the data size aggregating, eliminating redundant features or Cluster.





Data transformation $-2, 32, 100, 59 \rightarrow 0{\cdot}02, 0{\cdot}32, 1{\cdot}00, 0{\cdot}59$

Data Reduction

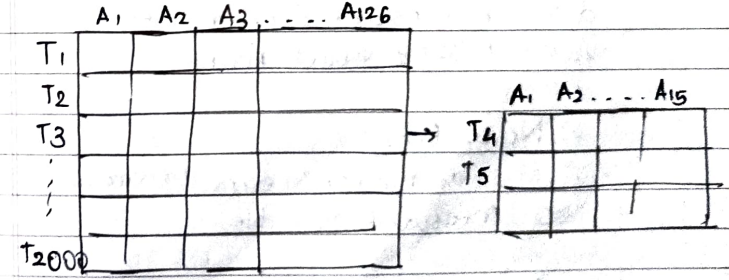


Fig Forms of Data Preprocessing.

Data Cleaning

1) Data Cleaning attends to fill the missing values, identifies the Outliers and, correct inconsistencies in data and smooth out the noise.

i) Missing Values

- Many tuples have no recorded value for several attributes such as cust_name

- Methods to fill missing values.

i) Ignore the tuples

ii) Use a global constant (unknow label $-\infty$)

iii) Use attribute mean to fill the missing values

iv) Use the most probable values for missing value. (Using Other customer attribute in data set consider decision tree to predict the missing value.

v) Fill missing values manually.

ii) Noisy Data

→ It's a random error or variance in a measured variable

a) - Data smoothing technique

1) Binning

Sorted data for price

4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equi depth) bins:

Bin1: 4, 8, 15

Bin2: 21, 21, 24

Bin3: 25, 28, 34

Smoothing by bin mean

Bin 1: 9, 9, 9.

2: 22, 22, 22

3: 29, 29, 29.

* Smoothing by bin boundaries

Bin 1: 4, 4, 15

2: 21, 21, 24

3: 25, 25, 34

b) Clustering

Outliers can be detected by clustering where similiar values are organized into groups or Clusters. Values that fall out of set of Clusters considered as outliers.

c) Regression
- Data can be smoothed by fitting data to a function.
Linear Regression finds best line to fit two variables, One variable can be used to predict other. Multiple Linear regression where more than 2 variables are involved.

→ Data Integration and Termination

Data Integration
Issues to Consider data integration
1) Schema integration → Entity identification
2) Redundancy
3) Detection and resolution of data value conflicts.

Data transformation: The data transformed or consolidated into form appropriate for mining. It involves:
1) Smoothing: Which work to remove voice from data.

2) Aggregation: Where summary or aggregation is applied to the data.

Generalization: Generalization of data where low level data replaced by higher level concepts. through use of hierachy.
Eg: Street → City or country
Age → young, middle age and senior

Normalization: Where attribute data are scaled to fall within a specified range
0.2-1

Attribute construction: Where new attributes are constructed and added from given set of attribute.

* Data Reduction:
- To obtain a reduced representation of the data set
- Mention integrity of original data.
1) Data cube aggregation
2) Dimension Reduction
3) Data compression.

Data Reduction

Data Cube Regreagation

Year = 1999
Year = 1998
Year = 1997

Quarter	Sales
Q1	224
Q2	408
Q3	350
Q4	158

Year	Sales
1997	156
1998	258
1999	359

Dimension Reduction: Dataset 100's of attributes many of which are irrelevant to mining task or redundant. Dimensionality reduces the size of dataset by removing such attributes or dimensions from it.

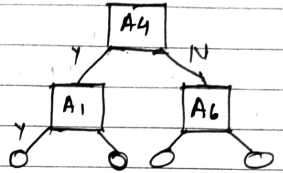
Forward Selection

Initial attribute set
[A1, A2, A3, A4, A5, A6]
Initial reduced set
{ }
→ {A1}
→ {A1, A4}
Reduced → {A1, A2, A6}

Backward Selection

Initial attribute set
{A1, A2, A3, A5, A6, A4}
⇒ {A1, A3, A4, A5, A6}
⇒ {A1, A4, A5, A6}
Reduced → {A1, A4, A6}

Decision Tree Induction



A1, A4, A6

Data Compression: Data encoding or transformation are applied to obtain reduced or compressed representation of data, if the original data is reconstructed from compressed data without any loss of info is called lossless

22.11.23

Unit - 2

Data Warehouse

6m

→ Provides tools for business executives to systematical organizes, understand use their data to make strategic decisions

6m ★ Key features of datawarehouse

1) Subject Oriented: It is organized around various subjects such as customers, product, suppliers and sales.

2) Integrated: Data warehouse is constructed by integrating multiple heterogenous sources.

3) Time variant - Data are stored to provide info from a historical perspective (past 5-10 yrs)

4) Non Volatile: Data warehouse is a physically separate store of data transformed from application data found in the Operational environment.

Difference betw Operational DB systems and datawarehouse.

— OLTP → cover day to day Operations
Eg: Banking, inventory, purchasing

— DW → Data analysis and decision making.

★12m	OLTP	OLAP
~~Characteristics~~ Features.		
~~Feats~~ 1) Characteristics	Operational processing	Informational processing
2) Orientation	Transaction	Analysis
3) Users	Clerk, DBA, DBP	Knowledged worker Eg manager, executive or analyst.
4) Function	Day to day operations	Long term informational requirements.
5) Database Design	ER based	Star or Snowflex, Subject Oriented

6) Access	Read and write	Mostly Read
7) No of records accessed	Tens (10)	Millions.
8) No of users	Thousands	Hundreds
9) DB size	100Mb - 1GB	100GB - 1TB.

- Data warehousing

Multidimensional Data Model

- View data in form of datacube
- Data cube --> Data to be modelled and viewed in multiple dimensions.

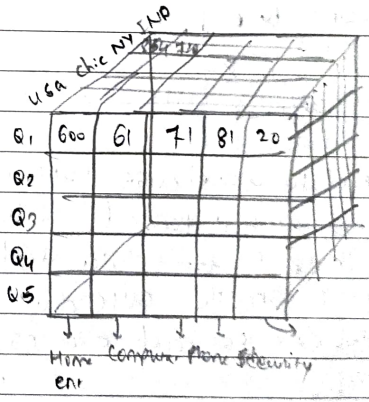
- Defined by dimensions by facts.

Eg: Dimensional → time, item, branch
Facts → Dollars_sold, units_sold

2-D view of Sales data

location = USA
item (type)

time	home ent	computer	phone	security
Q1	605	825	14	400
Q2				
Q3				
Q4.				



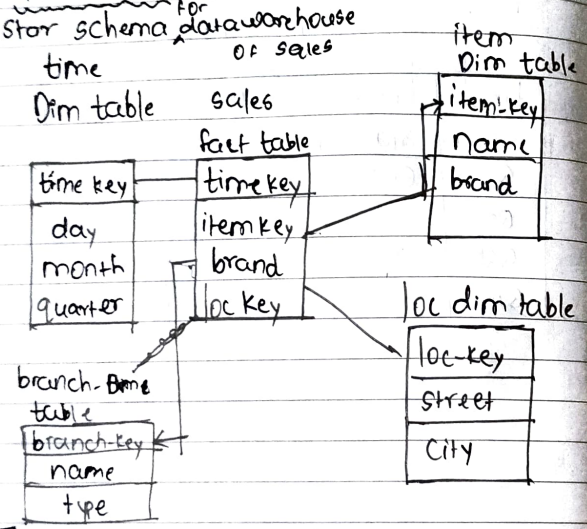
3-D Datacube representation.

12m-15m

Star, snowflake and fact constellations Schemas for multidimensional DB

1) Star schema

* Star schema for data warehouse of sales

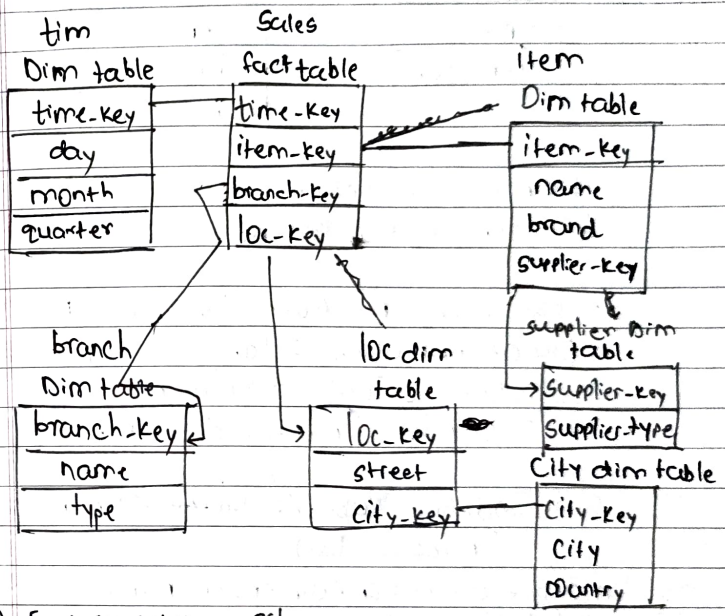


The data warehouse contains

a) A large central table (fact table) containing the bulk of data with no redundancy

b) A set of smaller attendant table (dimension table) one for each dimension.

2) Snowflake schema: It is the variant of star schema table where some dimension tables are normalized there by further splitting the data into additional tables.



3) Fact constolational schema

It requires multiple fact tables to share dimensional tables. This kind of schema can viewed as collection of stars. and hence it's called galaxy schema or fact constalation schema.

Sales fact table

time_key
item_key

Shipping fact table

item_key
shipping_key
time_key

Examples for defining star, snowflake and fact constellation schema

- SQL based on Query language - DMQL
- DW defined using 2 language primitives
  - One for Cube defination
  - One for dimension defination

Syntax: define cube ~~<Cubename>~~ (Cubename) <dimension list> : (measure list)

2) define dimension <dimension names> as (<attribute_list>)

Eg: Star schema

1) define cube Sales_star [time, item, branch, location]
   : dollars_sold = sum (sales_in_dollars), units_sold = count (*)

2) define dimension time as (time_key, day, month, yr)
   define dimension item as (item_key, name, branch)
   define dimension branch as (branch_key, name)

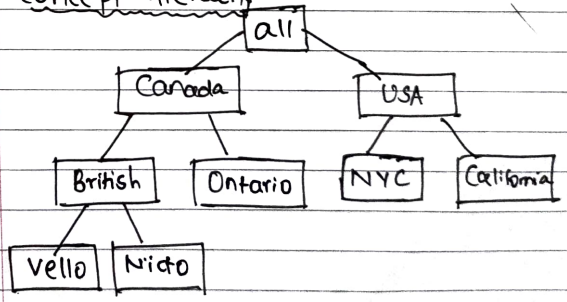
Eg: Snorflex schema

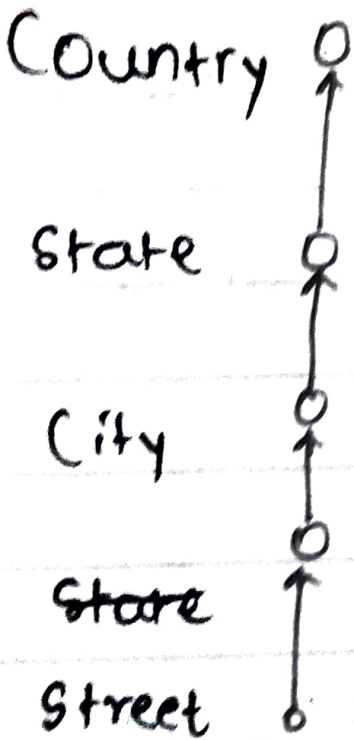
define Cubes ~~Snorflex~~ Sales_~~Snorflex~~ Star [time, item, branch]
: dollers_sold = sum (sales_in_dollars), units_sold = count (*)

define dimension time as (time_key, day, month, yr)
define dimension location as (~~location_key~~ City_key, city, month)
define dimension city as (city_name, cname, country)

* define dimension location as (location_key, street, city_key, state, country)
* define dimension item as (item_key, name, type, supplier (supplier_key, type)

* Concept Hierarchy





Defⁿ → Concept Hierachy defines a ~~sequence from~~ set of level of mapping from low concepts to higher level. The mapping forms a concept hierachy for the dimension locⁿ mapping a set of low level concepts (cities) to higher level concepts (country)