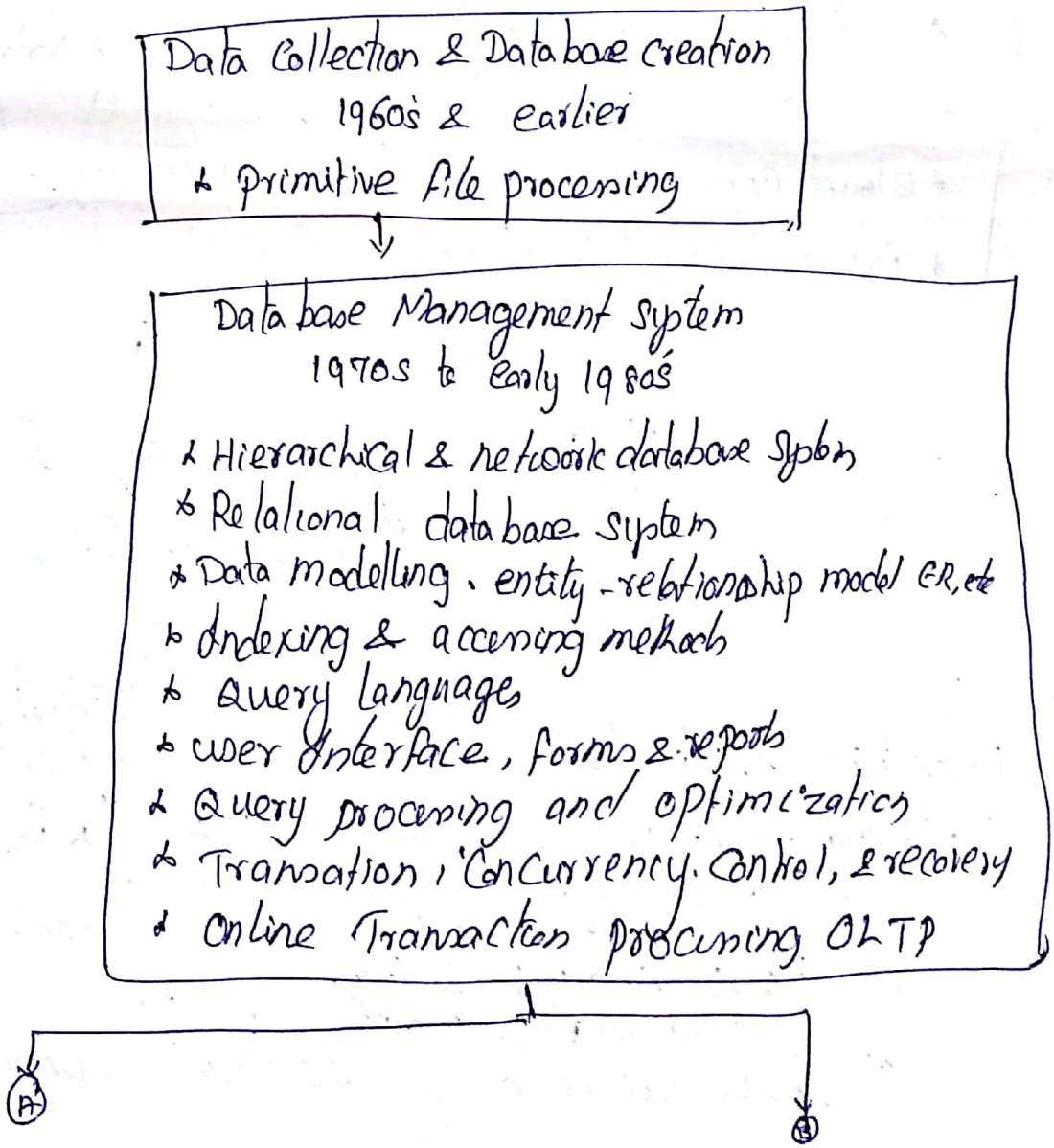
Unit III

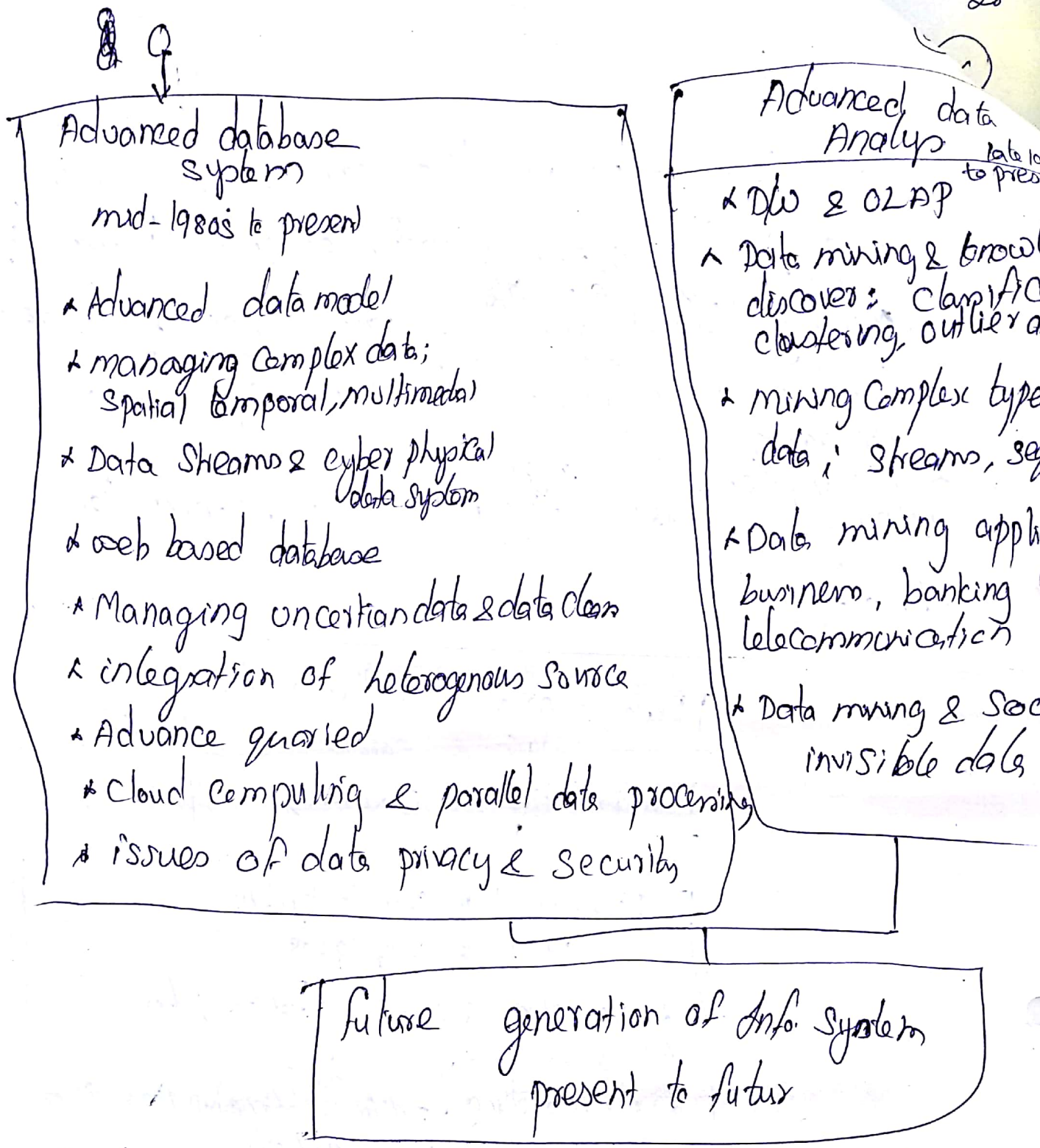
Data mining

- It is also called as knowledge discovery from data or KDD
- Extraction of interesting patterns or knowledge from huge amount of data
- Mining knowledge from data, knowledge extraction
- data/pattern analysis, data archaeology & data dredging
- Mining of knowledge or data.

(ii) Data mining as the Evolution of Information Technology.



Scanned by CamScanner



Applications

Potential Application

- Data Analysis & Decision Suppor
  - (i) Market Analysis & managemen
    Target marketing, Customer Relationship management (CRM)
  - (ii) Risk Analysis & Mgt
    - forecasting, quality control, etc
  - (iii) fraud detection & detection of unusual patterns (outli

Other Applications

- Text mining
- web mining
- Stream data mining
- ~~Pattern~~ Bioinformatics

Market Analysis & mgt

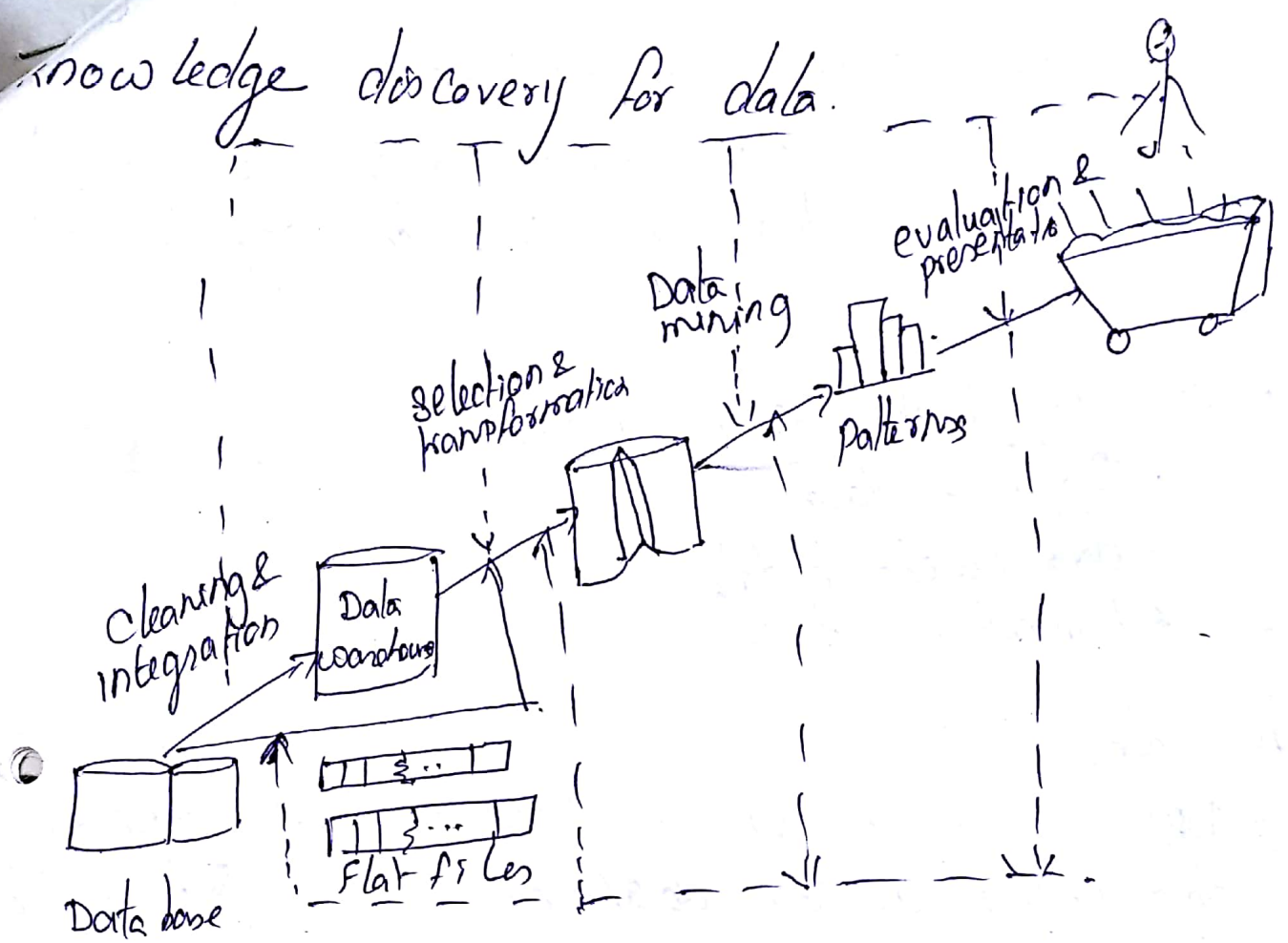
- Data Come from
  - Credit Card transactions, discount Coupons, Customer Complaint Calls et
- Target Marketing
  - find cluster of 'Model', Customer who share the characteristics such as interest, incomelevel etc
  - Determine customer purchasing patterns over time
- Cross Market analysis
  - Association / Corelations b/w product sales & prediction based on such associates

Fraud detection & mining unusual patterns

- Application: Healthcare, retail, Credit Card Service Co. telecomm.

Scanned by CamScanner

Knowledge discovery for data.



Data cleaning : Remove noise & inconsistent data

Data integration : Multiple data source may be combined

Data selection : data relevant to the analysis task are retrieved from the database

Data transformation : data are transformed & consolidated into forms by performing summary or aggregation operations

Data mining : Intelligent methods are applied to extract data patterns

Pattern evaluation : to identify the truly interesting patterns representing knowledge based on interesting measures

Scanned by CamScanner

Knowledge presentation: where visulaization & knowledge representation techniques are used to present mined knowledge to users

What kind of Data Can be mined

- Database data
- data warehouse data
- transactional data

Database data

- DBMS Consists of a Collection of interrelated data known as as database, and a set of s/w pgms to manage and acces the data.
- s/w pgm provide mechanisms for database structures and data storage for specifying & managing Concurrent, shared and distributed data accen.

- A Relational database is a collection of tables each of which is assigned a unique name.
- Each table Consist of a set of attributes (Columns or fields) and stores a large set of tuples (records or
- Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values

Scanned by CamScanner

A Semantic data model such as Entity relationship (ER) data model is constructed for relational data base

* ER model represents the database as a set of entities & their relationships

eg. A relational database for AllElectronics.
It Consist of following relational tables
customer, item, employee & branch.

The relational table Customer Consist of set of attributes describing the customer information including
customer cust_id, name, address, age, occupation, annual_income, etc.

item - item_id, brand, category, type, price, place_made, supplier, cost.
employe - emp_id, name category, group, salary....
branch - branch_id, name, address...

The tables can be used to represent the relationships b/w or among multiple entities: eg,

purchases - Customer purchase items, creating a sales transaction handled by an employee.
item sold - list of item sold in a given transaction
work_at - employee work at a branch of Allelectron

Scanned by CamScanner

(6)

Relational schema for a relational database.

customer (cust-ID, name, address, age, ...)
item (item-id, brand, Category, type, price...)
employee (empl-ID, name, Category, group, salary...)
branch (branch-ID, name, address, ...)
purchases (trans-id, cust-ID, empl-ID, date, time, method_paid, amount)
item_sold (trans-id, item-ID, qty)
works_at (empl-id, branch-id)

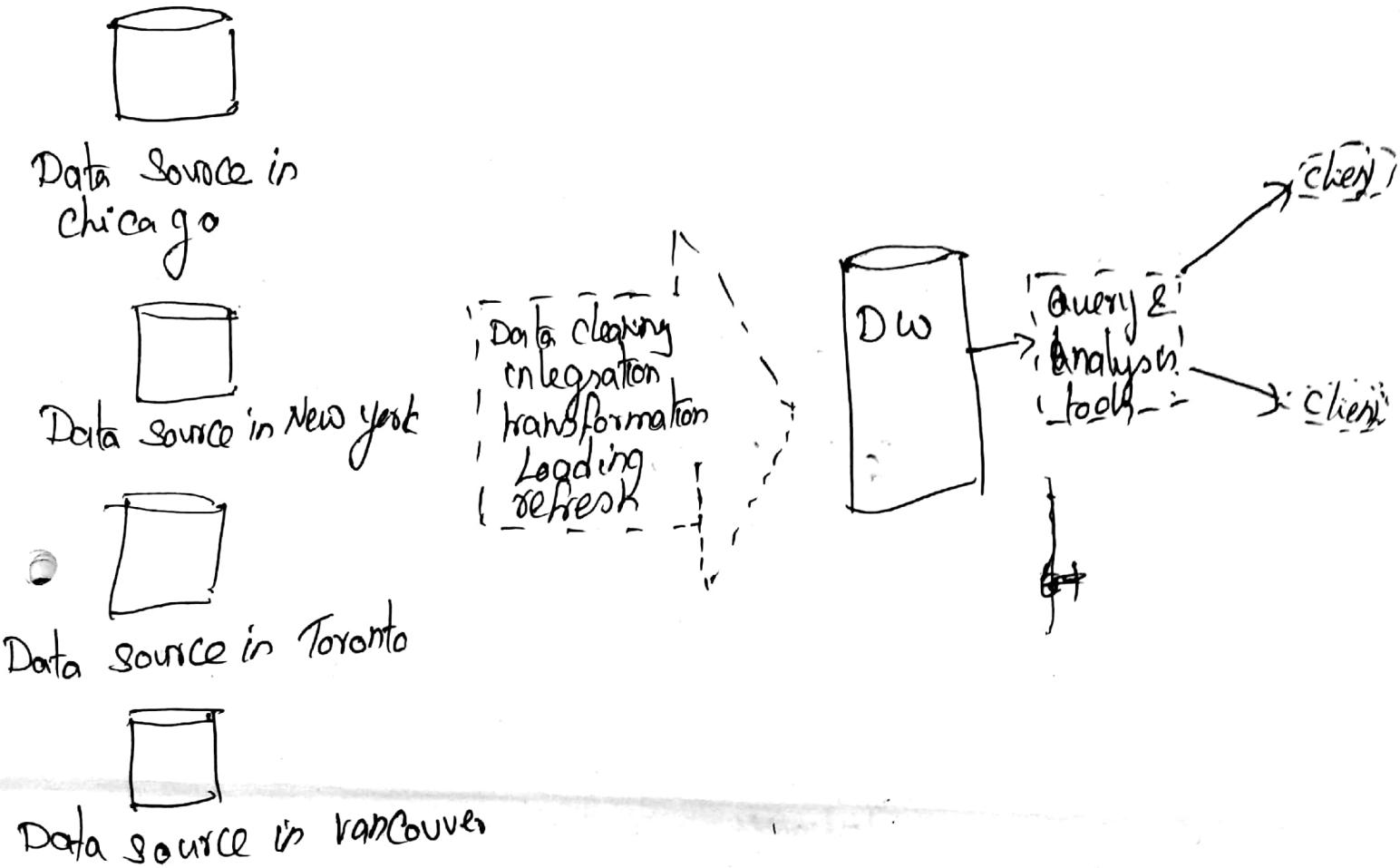
Relational data can be accessed by using database query written in relational query language. eg selection, join

Data warehouses:

- is a respository of information collected from multiple sources. Stored under a unified schema and residing at a single site.
- Data warehouses are constructed via a process of data cleaning, data integration, data transformation data loading & data refreshing.

Scanned by CamScanner

typical frame work for All Electronics



- A data warehouse is modeled by a multidimensional data structure called a data cube.
- in which each dimension correspond to ~~set of~~ an attribute or a set of attributes in the schema.
- each cell stores the value of some aggregation measures count or sum.
- A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data

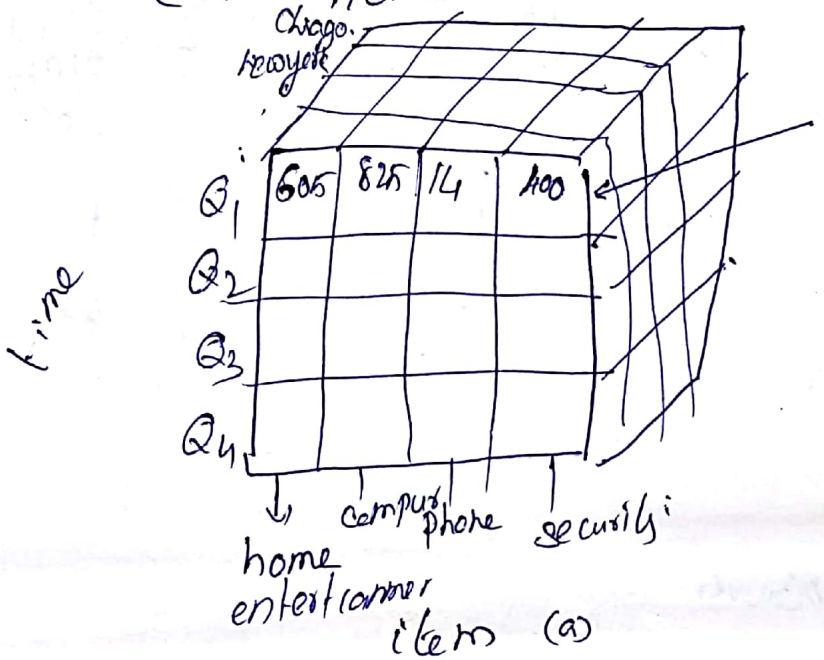
Scanned by CamScanner

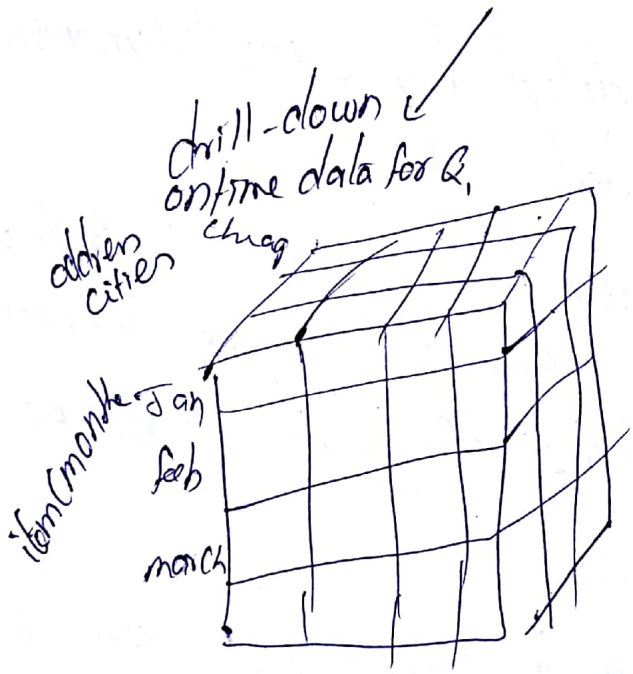
eg. A data cube for All electronics

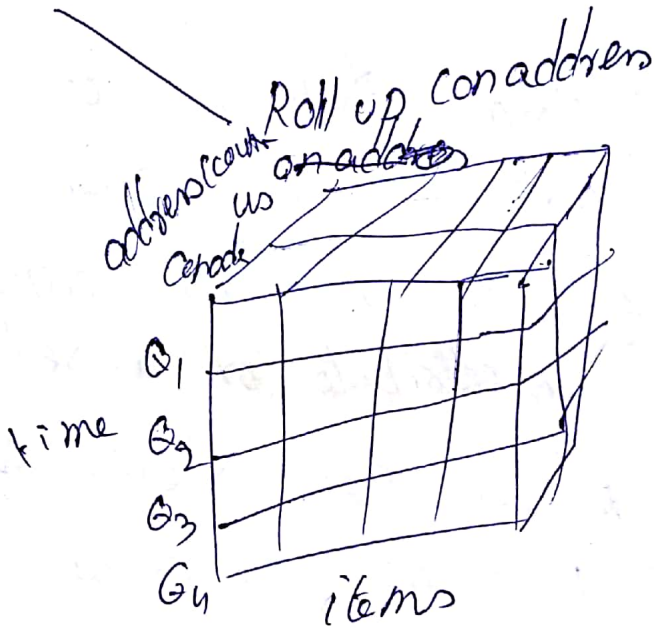
- Cube has three dimensions. (i) addres (with city value Chicago, Newyork, Toronto, vancouver)

(ii) time (with Quarter values Q_1, Q_2, Q_3, Q_4)

(iii) item (values home entertainment, Computer, Phone Security







Scanned by CamScanner

Transactional data:

- each record in a transactional database captures a transaction such as customer's purchase etc.
- transaction includes a unique transaction identity number and list of the items making up the transactions.

eg

Trans id	list-of-item-ids
T100	I_1, I_3, I_8, I_{16}
T_{200}	I_2, I_8
⋮	⋮

Other kind of Data.

- time related or sequence data
  (eg) historical data records, stock exchange data, etc.
- data streams:
  (eg) video surveillance & sensor data, etc
- spatial data
  eg. maps.
- engineering design data
  - design of buildings, system components etc
- hypertext & Multimedia data.
  (eg) text, image, video & Audio data
- graph & N/w data
  - social & info N/w data
- web
  - huge & widely distributed data

Scanned by CamScanner

Data mining functionalities: or what kinds of patterns can be mined

- It is used to specify the kind of pattern to be found in data mining task.
- Data mining task are classified into two categories
  (i) descriptive
  (ii) predictive.

- Descriptive data mining task
  - it characterize the general properties of data in the database

Predictive data mining task
- it perform inference on the current data in order to make predictions

Functionalities
(i) Concept/class description: characterization and discrimination
- Data can be associated with classes or concepts.
  eg. In All electronic store classes of item for sale include computers & printers
  Concepts include big spenders or budget spenders.

Scanned by CamScanner

(11)

data characterization,

- by summarizing the data of the class under study (target class)

data discrimination

- by comparison of the target class with one or a set of comparative classes (contrasting class)

Data characterization.

- is a summarization of the general characteristics or features of a target class of data.
- the data cube based OLAP rollup operation can be used to perform user-controlled data summarization
- An attribute-oriented induction technique can be used to perform
- the o/p of data characterization can be of various forms

  eg. pie charts, bar charts, curves, multidimensional data cubes & multidimensional tables

Scanned by CamScanner

- Data discrimination.
  - It is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting class
  - It is expressed in the form of rules is called discriminant rule

Mining Frequent patterns, Associations and Correlation

Frequent patterns:
- patterns that occur frequently in data.
- Many kinds of frequent patterns like, frequent itemsets, frequent subsequence, frequent substructure
- Frequent itemset:
  it refers to a set of items that often appear together in a transactional data set (eg) Milk etc.
- frequent subsequency
  - the patterns that the customers tend to purchase

  eg laptop → digital camer — memory card

Scanned by CamScanner

- Substructure refers different structural forms eg. graphs, trees or lattices. they may combined with itemset or subsequence
- if the substructure occur frequently it is called (frequent) structured patterns.

- Mining frequent patterns leads to the discovery of interesting associations and correlation within data

eg. Association Analysis

eg If the manager wants to know which items are frequently purchased together i.e within same transaction

buys (X, "computer") ⟹ buys (X, "software") [support = 1%, confidence = 50%]

where X is a variable representing Customer.

- A Confidence or Certainity of 50% means that if a Customer buys a Computer there is a 50% chance that Customer will buy S/w also.
- A 1% support means that 1% transaction under analysis show that Computer & S/w purchase together
- Here association rule involves a single attribute

for association rule X→Y, the support of the rule is denoted as Sup(X→Y) and is no. of transaction where XUY appears divided by the total number of transaction.

Confidence of the rule XUY appear divided by the number of transaction where X appears

Scanned by CamScanner

or predicate ie buys that repeat..

* Association rule that Contain a single attribute or predicate are referred as single dimensional anocation rule

eg. age(X, "20..29") ∧ income(X, "40k..49k") ⇒ buys (X, "laptop") [support = 2%, Confidence 60%

2% of 20 to 29 year old customer with income 40k to 49k have purchased laptop.

60% probability that a Customer in this age and income group will purchase a laptop.

* More than one attribute or predicate is used ie age, income, buy. So

* More than one attribute is referred as Multidimensional anociation rule

3. Classification & Regression for predictive Analysis

Classification:

* is the procen of finding a model or function that describes and distinguishes data Clanses or Concepts.

* the model are derived based on the analysis of a set of training data

Lift = $\frac{(A+B)}{\frac{\text{Total}}{B_{\text{Total}}}}$

(Sub) Support = $\frac{(A+B)}{\text{Total}}$ Confidence = $\frac{A+B}{B}$

Scanned by CamScanner

(17)

ie data object for which the class labels are known.

- The derived model may be represented in various forms such as classification rules ie IF-THEN, decision trees, Mathematical formulae or neural Nws

A decision tree

- is a flow chart like tree structure where each node denotes a test on an attribute value, each branch represent an outcome of the test and tree leaves represents classes or classification
- it can be converted to classification rules

Neural Nw:

used for classification, is a collection of neuron like processing units with weighted connections between the units

Regression

- are continuous valued functions
- ie, it is used to predict missing or unavailable numerical data values rather

Scanned by CamScanner

— than class label prediction.

< CE is used for numeric prediction

If then.

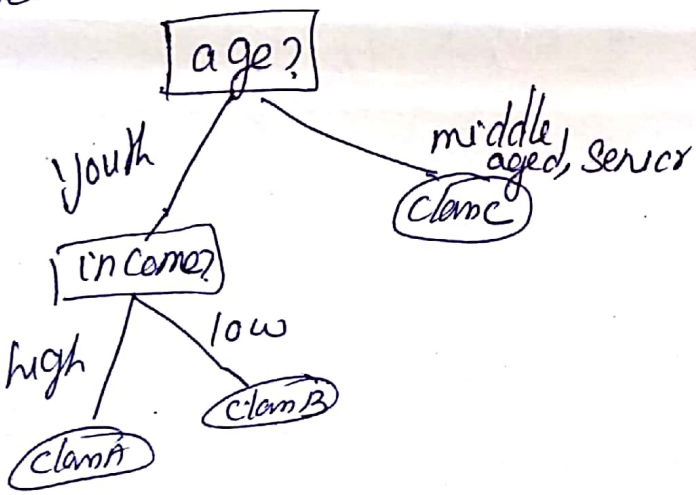
age (x, "youth") And income (x, "high") ⟶ class(x, "A")

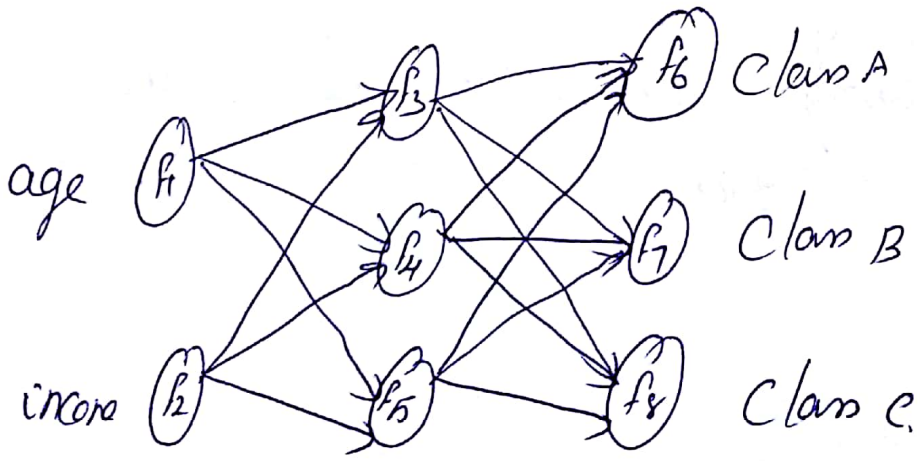
age (x, "youth") And income (x, "low") ⟶ class(x, "B")

age (x, "middle_aged") ⟶ class (x, "C")

age (x, "senior" ⟶ class (x, "C")

decision tree

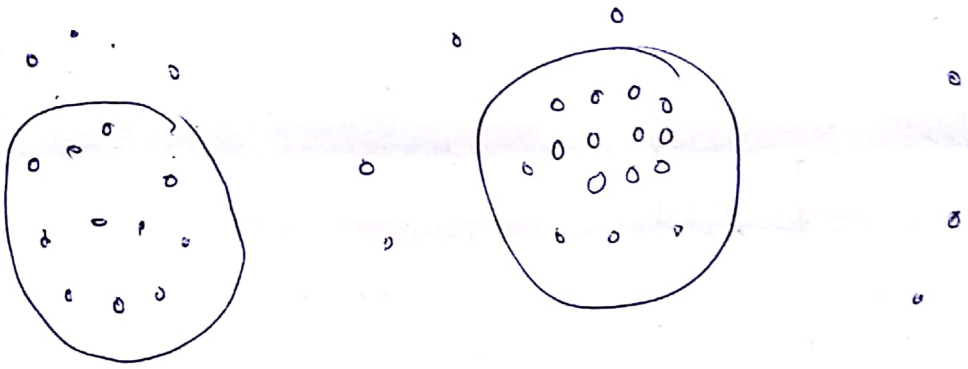




Scanned by CamScanner

4. Cluster Analysis

- Clustering is used to generate class labels for a group of data.
- Objects are clustered or grouped based on the principle of maximizing the intraclass similarity & minimizing the interclass similarity.

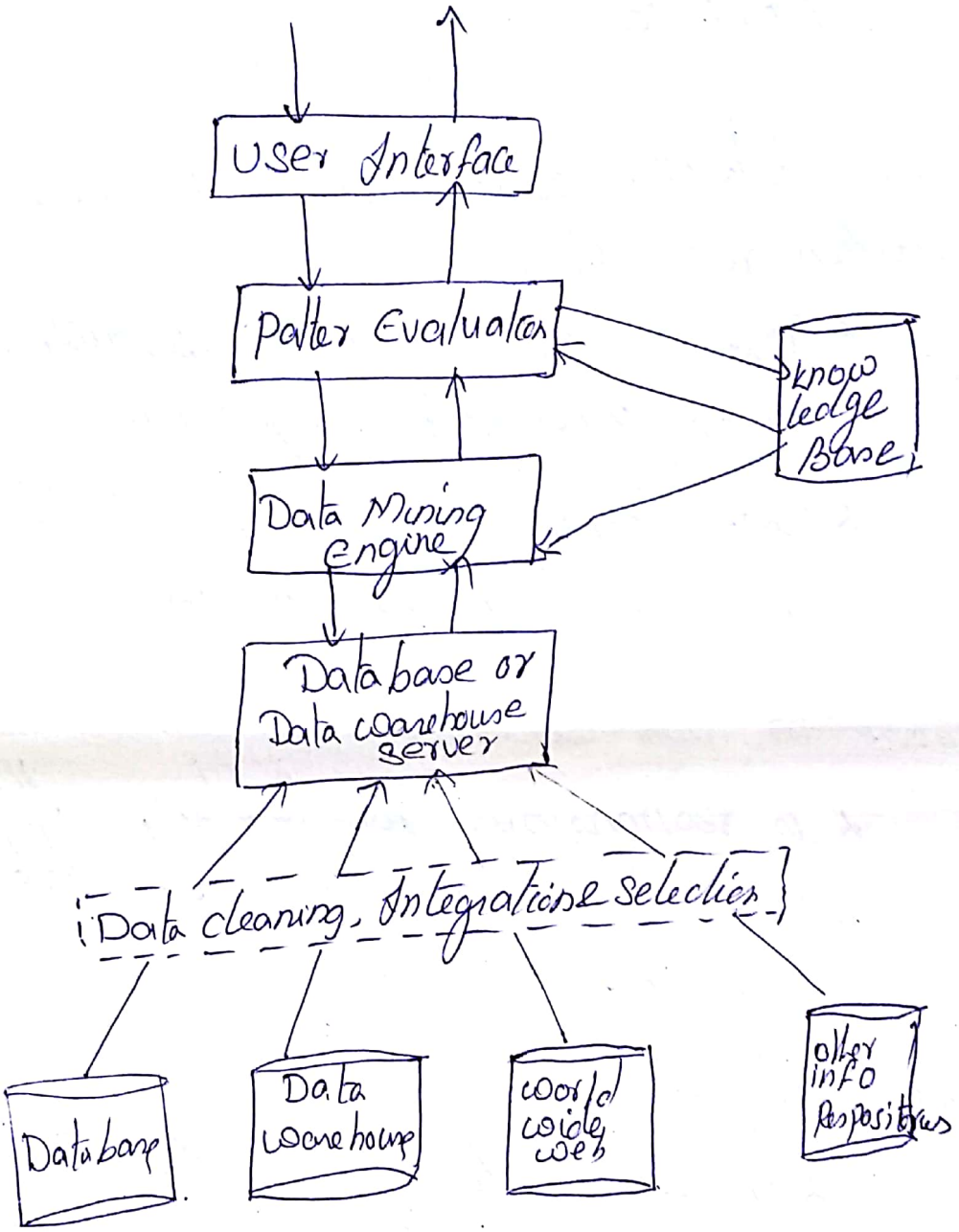


5) Outlier Analysis

- Data set may contain objects that do not comply with the general behavior or model of data
- The analysis of outlier data is referred to as outlier Analysis or anomaly mining

Scanned by CamScanner

Architecture of Typical Data Mining Systems



Components are

- Database, data warehouse, WWW or other information Repository
- Database or Data warehouse server
- Knowledge base
- Data mining Engine

Scanned by CamScanner

- Pattern Evaluation Module
- User Interface

(i) Database, data warehouse, World wide web, or other information repository:

- ~~This~~ is one or set of db, dwh, www or other information repository
- Data cleaning, integration selection techniques may be performed on the data

(ii) Database or data warehouse Server:

- is responsible for fetching the relevant data, based on the user's data mining request.

(iii) knowledge base:

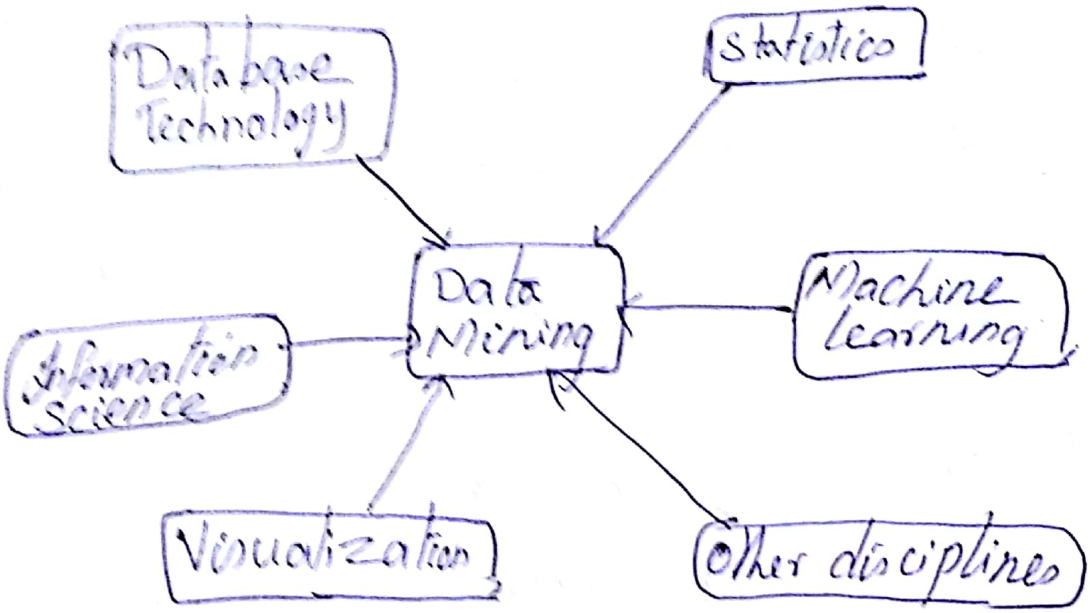
- is used to guide the search or evaluate the interestingness of resulting pattern. ~~Such knowledge can include~~
- ~~Concept hierarchies etc~~
- knowledge can include concept hierarchies used to organize attributes or attribute values into different levels

Scanned by CamScanner

* Data mining Engine:
  * Set of functional Modules for task such as Characterization, association and Correlation analysis, classification, prediction, cluster analysis, outlier analysis & evolution analysis
* Pattern evolution Module:
  * employ interesting new measures and interact with the data mining modules so as to focus the search towards interesting patterns.
  * Pattern Evolution module is integrated with the mining module, depending on the implementation of the data mining method used
* User Interface:
  * It Communicates b/w users and the data mining system along the user to interact with the system by specifying a data mining Query or task ~~providing~~,

Scanned by CamScanner

Classification of Data Mining System



- DM Consist of Set of disciplines including
  - database Systems.
  - Statistics.
  - M/c learning
  - visualization
  - information scene
  - other disciplines include Neural N/ws, fuzzy and/or rough set Theory, knowledge representation, inductive logic programming or high-performance Computing

Scanned by CamScanner

classification according to the kind of database mined.

- data mining system can be classified according to the kinds of database mined.
- Database systems can be classified according to criteria such as data models or types of data or application involved.
- if classifying according to data model then we have relational, transactional, object-relational, or data warehouse mining system.
- If according to types of data, then spatial, time series, text, stream data, Multimedia or we have mining system

Classification according to the kinds of knowledge Mined.

- Based on the data mining functionalities such as characterization, discrimination, association and correlation analysis, classification, prediction, clustering, Outlier and evolution analysis

Scanned by CamScanner

Classification according to the kinds of techniques utilized:

- the techniques can be described according to the degree of user interaction involved (autonomous systems, interactive, exploratory, query driven) or methods of data analysis employed (database oriented or data-oriented techniques, M/c learning)

Classification according to the application adopted.

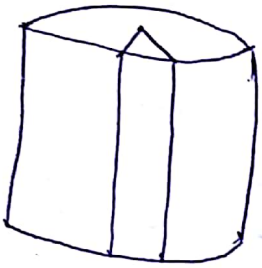
- DM can be categorized according to the applications they adapt.
  eg. for finance, telecommunications, DNA, stock markets, e-mail...

Data mining task Primitives

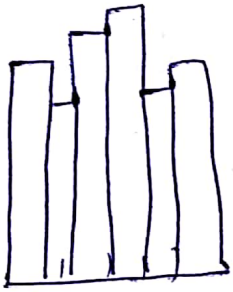
- Data mining task can be specified in the form of a data mining Query. which is the input to the data mining system.
- Data mining Query is defined in terms of data mining primitives.
- primitives allow the user to interactively communicate with the data mining system during discovery of data

Scanned by CamScanner

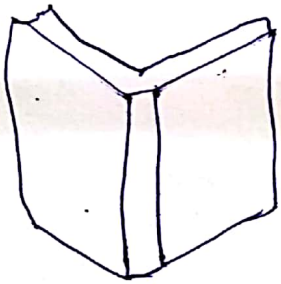
The data mining primitives Specify the following.



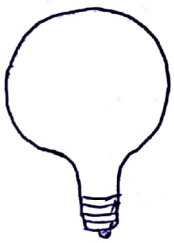
Task-relevant data
- Database or data warehouse name
- Database tables or data warehouse cubes
- Conditions for data selection
- Relevant attributes or dimensions
- Data grouping Criteria



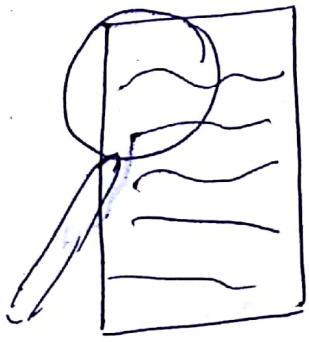
knowledge type to be mined
- Characterization
- Discrimination
- Association/Correlation
- Classification/ prediction
- clustering



background knowledge
- Concept hierarchies
- user beliefs about relationships in the data



Pattern interestingness measures
- Simplicity
- Certainity (eg Confidence)
- Utility (eg support)
- Novelty



Visualization of discovered patterns
- Rules, tables, reports, charts, graphs, decision trees, & cubes
- Drill-down & rollup

Scanned by CamScanner

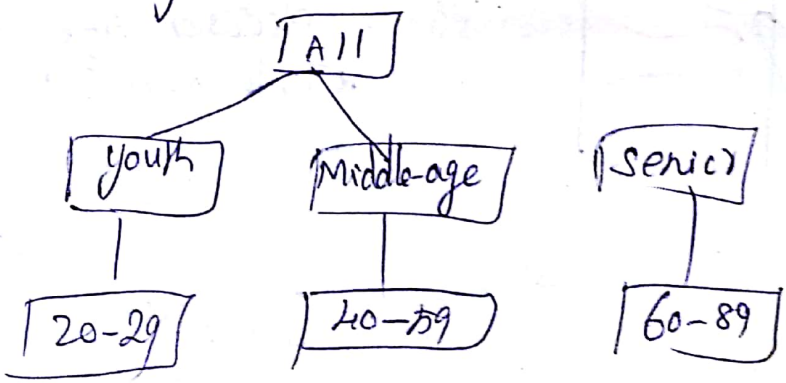
Data mining primitives

- The set of task relevant data to be mined
- The kind of knowledge to be mined
- The background knowledge to be used in the discovery process
- The interestingness measures and thresholds for pattern evaluation.
- The expected representation for visualizing the discovered patterns.

Data Mining Language [DMQL]

- The design of an effective data mining Query Language requires a deep understanding of the power, limitation and underlying mechanisms of the various kinds of data mining task

(eg) Concept hierarchy for attribute (or dimension) age



Scanned by CamScanner

Major issues of Data Mining

(i) Mining Methodology & User Interaction issues
(ii) ~~Performance issues~~ (iii) Efficiency & Scalability
(iii) diversity of database type issues
(v) Data Mining & Society

(i) Mining Methodology & User Interaction issues:

Mining Methodology issues

- Mining different kinds of knowledge in database:
  - users can be interested in different kinds of knowledge, data mining cover world wide spectrum of data analysis and knowledge discovery tasks including data characterization, discrimination, association and Correlation analysis, classification, prediction, clustering, outlier analysis
  - ~~Interactive~~
- Mining knowledge in Multidimensional Space
  - Can search for interesting patterns among Combinations of dimensions (attributes) at various levels of abstraction
  - Such mining is known as (exploratory) Multidimensional data mining.
  - Mining knowledge in cube space Can enhance the power & flexibility of data mining

Scanned by CamScanner

Data Mining - a interdisciplinary effort:

- The power of data mining can be substantially enhanced by integrating new methods from multiple disciplines.
  - eg (i) mine data with natural language text
  - (ii) The mining of S/w bugs in large programs

Boosting the power of discovery in Networked environment

- data objects reside in a linked or interconnected environment, it may be web, database relations, files or documents
- knowledge derived in one set of objects can be used to boos the discovery of knowledge in a related or semantically linked set of objects

- Handling of Data
  - uncertainty, noise or incompleteness
  - * Data contain noise, error, exceptions etc.
  - Errors and noise may confuse the data mining process leading to the derivation of erroneous pattern.

Scanned by CamScanner

* Data cleaning, data preprocessing, outlier are examples of techniques need to integrated with data mining proces

* Pattern Evaluation and pattern or Constraint-guided Mining:

  By using interesting ness measures or user specified Constraints to guides the discovery proces.

User Interaction issues

* Interactive Mining
  * DM is highly interactive
  - Interactive mining should allow users to dynamically change the focus of a search, to refine mining request based on return results and to drill, dice and pivot thro the data and knowledge space interactively

* Incorporation of background knowledge:
  * Background knowledge, Constraints, rules and other information regarding the domain should be incorporated into the knowledge discovery proces

Scanned by CamScanner

Adhoc data mining and DMQL:

- Query languages have played an important role in flexible searching because the allow users to pose adhoc queries
- High level DMQL or highlevel flexible user interfaces will give users to define adhoc data mining tasks

Presentation & visualization of data mining results:

- How to a data mining system present data mining results, vividly & flexibly. So that the discovered knowledge can be easily understood and usable by humans.

3. Efficiency & Scalability:

- Efficiency & Scalability of data Mining algorithms
  - Data Mining - algorithms must be efficient & scable in order to effectively extract information from huge amounts of data in many data respositories.
- Efficiency, Scalability, performance, optimization and the ability to execute in real time are key criteria for the development of new data mining algorithms

Scanned by CamScanner

- parallel, distributed and incremental Mining algorithms
  - Size of data set, distribution of data & Computational Complexity are factors that motivate the development of parallel & distributed data-intensive mining Algorithms
  - Algorithms first partition the data into piecer. Each piece is process in itself by searching for patterns
  - Cloud Computing & cluster Computing use Computers on a distributed and collobrative way to solve very large-scale Computational task

(4) Diversity of Database type.
- Handling Complex types of data
  - it generate wide spectrum of new data types, from structured data such as relational & data warehouse data to semi structured and un-structured data: from stable data respositories to dynamic data from simple data to temporal data object etc.

Scanned by CamScanner

* Mining dynamic, networked and global data respositories:
  * Multiple Source of data are connected by the internet & various kinds of N/w forming distribute & heterogenous global information systems & NWs.

5. Data Mining & society
  * Social impacts of data Mining:
  * Based on the everyday lives it is important to study the impact of data mining in Society.
  * privacy - preserving data mining:
    * DM will help scientific discovery, busines management, economy and Security protection.
  * Invisible data Mining:
    * Intelligent Search engines & Internet based Stores perform such invisible data mining by incorporating data mining into their Components to improve their functionality & performance

Scanned by CamScanner