CMR
INSTITUTE OF
TECHNOLOGY

Internal Assesment Test – I- Solution

SUB Data Warehousing & Data Mining Code: 13MCA442

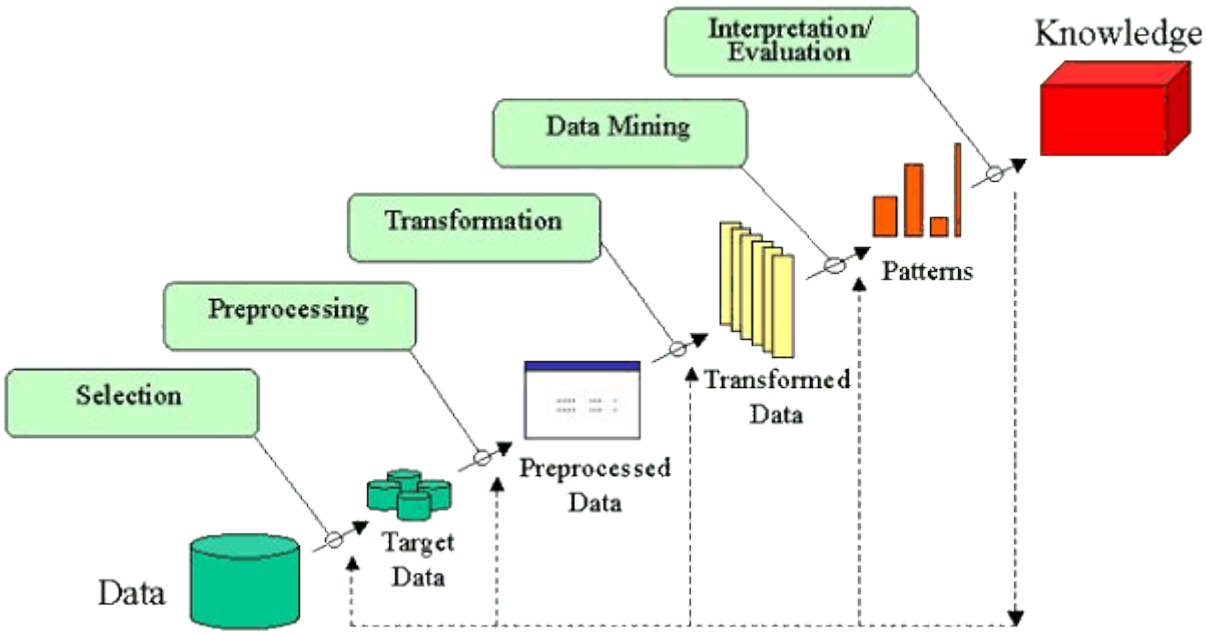
1 a. Explain Knowledge discovery process in data mining.

Ans The term *Knowledge Discovery in Databases*, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.

The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, subsampling, and transformations of that database.

An Outline of the Steps of the KDD Process



1 b Explain any two data preprocessing steps.

Ans: **Aggregation: Combining two or more attributes (or objects) into a single attribute (or object)**

Purpose

- **Data reduction**
  - **Reduce the number of attributes or objects**
- **Change of scale**

- **Cities aggregated into regions, states, countries, etc**
- **More "stable" data**
  - **Aggregated data tends to have less variability**

Sampling: Sampling is the main technique employed for data selection.
- **It is often used for both the preliminary investigation of the data and the final data analysis.**

Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.

Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

Types of Sampling

Simple Random Sampling
- There is an equal probability of selecting any particular item
- Two types

Sampling without replacement
- As each item is selected, it is removed from the population

Sampling with replacement
- Objects are not removed from the population as they are selected for the sample.
  - In sampling with replacement, the same object can be picked up more than once

Stratified sampling
- Split the data into several partitions; then draw random samples from each partition

2a. Explain different type of attributes and data sets.
- Ans: There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Types of data sets

- **Record**
  - **Data Matrix**
  - **Document Data**
  - **Transaction Data**
- **Graph**
  - **World Wide Web**
  - **Molecular Structures**
- **Ordered**
  - **Spatial Data**
  - **Temporal Data**
  - **Sequential Data**
  - **Genetic Sequence Data**

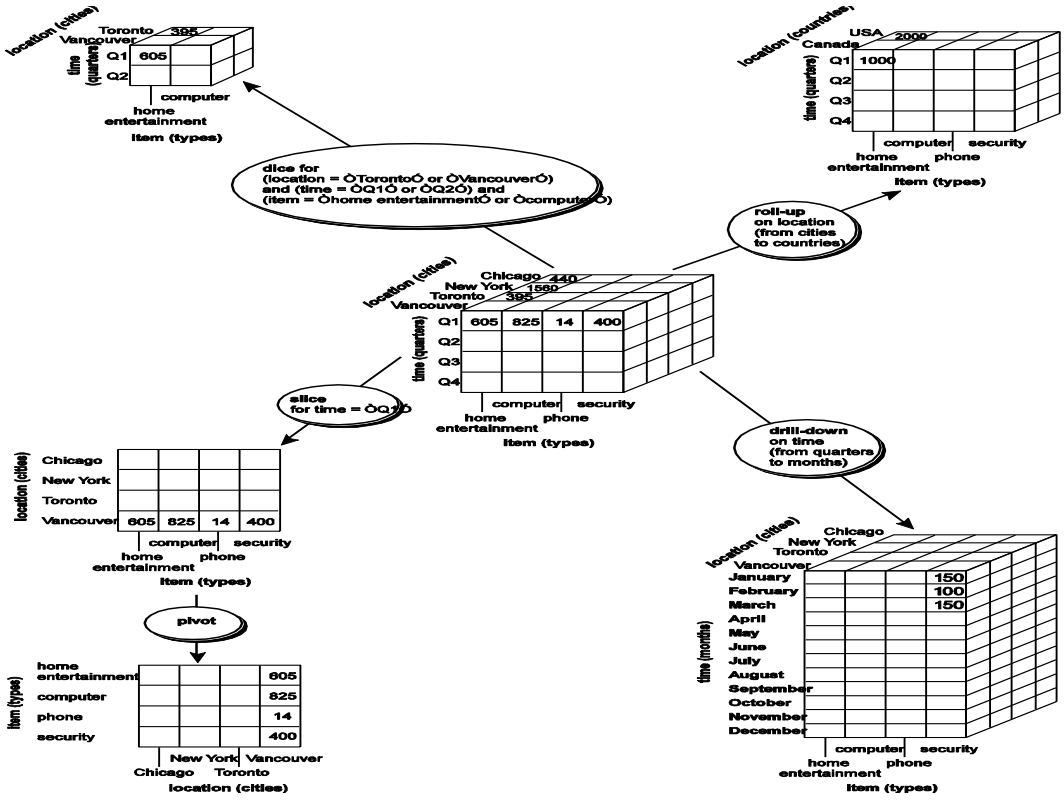
3a. Explain the type of operation that can be performed on data cube.

Ans: A number of operations may be applied to data cubes. The common ones are:

- roll-up (increasing the level of abstraction)
- drill-down (increasing detail)
- slice and dice (selection and projection)
- pivot (re-orienting the view)

- *Roll-up* (less detail) - when we wish further abstraction (i.e. less detail). This operation performs further aggregation on the data, for example, from single degree programs to Schools, single countries to Continents or from three dimensions to two dimensions.
- *Drill-down* (increasing detail) - reverse of roll up, when we wish to partition more finely or want to focus on some particular values of certain dimensions. Drill-down adds more detail to the data, it may involve adding another dimension.

- *Slice and dice* (selection and projection) - the slice operation performs a selection on one dimension of the cube (e.g. degree = "MIT"). The dice operation performs a selection on two or more dimensions (e.g. degree = "BIT" and country = "Australia" or "India")
- *Pivot* (re-orienting the view) - an alternate presentation of the data e.g. rotating the axes in a 3-D cube.



4a. Explain the difference between ROLAP and MOLAP.

Ans: Relational OLAP (ROLAP)

- Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware to support missing pieces
- Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
- greater scalability

- Multidimensional OLAP (MOLAP)
  - Array-based multidimensional storage engine (sparse matrix techniques)
  - fast indexing to pre-computed summarized data

4b. What are the issues related to proximity measures.

Ans: (i) How to handle the case in which attributes have different scales and /or correlated.
(ii) how to calculate proximity between objects that are composed of different types of attributes.
(iii) how to handle proximity calculation when attribute have different weights.

Q5a. List the major steps involved in the ETL process.

- ANS: ETL process consists of data extraction and data transformation.
- Data transformation data cleaning and data loading.
- Data cleaning deals with detecting and removing errors and inconsistencies from the data.

- Building an integrated database from a number of source systems may involve solving some or all of the following problems:

(i) Instance identity problem: same customer or client may be represented slightly differently in different source systems.

(ii) Data Errors: Many different types of data errors other than identity errors are possible:

- Missing values.
- Meaning of some code values may not be known.
- Duplicate records.
- Wrong aggregations.
- Inconsistent use of nulls, spaces and empty values.
- Some attribute values may be inconsistent(outside their domain)
- Data may be wrong because of input errors.

(iii) Record linkage problem: relates to the problem of linking information from different databases that relates to the same customer.

(iv) Semantic integration problem: deals with integration of information found in heterogeneous OLTP and legacy sources. Some of the sources may be relational, some may not be.

(v) Data integrity problem: This deals with issues like referential integrity, null values, domain of values, etc.

Q5b Explain FASMI characteristics of OLAP.

Ans:

FASMI Characteristics

- *Fast*: OLAP queries should be answered very quickly, perhaps within seconds.
- *Analytic*: An OLAP system must provide rich analytic functionality and it is expected that most OLAP queries can be answered without any programming.
- *Shared*: An OLAP system is a shared resource although it is likely to be accessed only by a select group of managers. Being a shared system, an OLAP system should provide adequate security for confidentiality as well as integrity.
- *Multidimensional*: It must provide a multidimensional conceptual view of the data. A dimension often has hierarchies that show parent/child relationships between the members of a dimension. The multidimensional structure should allow such hierarchies.
- *Information*: OLAP system usually obtain information from a data warehouse. The system should be able to handle a large amount of input data.

Q6: What is a data ware house? How we can implement a data ware house.

Ans: According to W. H. Inmon: *A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process.*

Important to note subject-oriented, integrated, and time-variant properties of a data warehouse.

Subject-oriented

- A DW is organized around major subjects, such as student, degree, country.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations.
- A DW provides a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Integrated

- A DW may be constructed by integrating information from multiple data sources e.g. multiple OLTP databases.
- Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources.

Time Variant

- A DW usually has long time horizon, significantly longer than that of operational systems.
    - Operational database: current value data.
    - DW data: provide information from a historical perspective (e.g. past 5-10 years)

- Every key structure in the DW contains an element of time, explicitly or implicitly
- Operational data may or may not contain time element.

Non-volatile

- A physically separate store of data transformed from the operational environment.
- No update of data
- Does not require transaction processing, recovery, and concurrency control mechanisms
- Requires only two operations in data accessing: *initial loading of data* and *access of data*.

Implementation Steps:

I. *Requirements analysis and capacity planning*: defining enterprise needs, defining architecture, carrying out capacity planning and selecting the H/W and S/W tools.
II. *Hardware integration*: Integrating the servers, the storage devices and the client software tools.
III. *Modeling*: Designing the warehouse schema and views.
IV. *Physical Modeling*: Designing the physical data warehouse organization, data placement, data partitioning, deciding on access methods and indexing.
V. (V) *Sources*: Identifying and connecting the sources using gateways, ODBC drivers or other wrappers.
VI. (VI) *ETL*: ETL process may involve identifying a suitable ETL tool vendor and purchasing and implementing the tool. This may also include customizing the tool to suit the needs of the enterprise.
VII. *Populate the data warehouse*: Testing the tool using staging area. Populating the warehouse using ETL tool.
VIII. (VIII) *User applications*: Designing and implementing applications required by the end users.
IX. (IX) *Roll-out the warehouse and applications*: The warehouse system and the applications may be rolled out for the user community to use.

Q7a: Explain Similarity and Dissimilarity. What is metric?

- Similarity
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

- Distances, such as the Euclidean distance, have some well known properties.
    1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
    2. $d(p, q) = d(q, p)$ for all p and q. (Symmetry)
    3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r. (Triangle Inequality)

  where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q.
- A distance that satisfies these properties is a metric

Q7b: If p and q are two given data objects
p = 1 0 0 0 0 0 0 0 0 0
q = 0 0 0 0 0 0 1 0 0 1
Calculate the SMC and Jaccard Coefficients.

Ans: p = 1 0 0 0 0 0 0 0 0 0
q = 0 0 0 0 0 0 1 0 0 1
M_{01} = 2 (the number of attributes where p was 0 and q was 1)
M_{10} = 1 (the number of attributes where p was 1 and q was 0)
M_{00} = 7 (the number of attributes where p was 0 and q was 0)
M_{11} = 0 (the number of attributes where p was 1 and q was 1)

SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ = (0+7) / (2+1+0+7) = 0.7
J = (M_{11}) / $(M_{01} + M_{10} + M_{11})$ = 0 / (2 + 1 + 0) = 0

Q8: Write the difference between OLAP and OLTP.

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	Dozens
DB size	100MB-GB	100GB-TB