B.E.

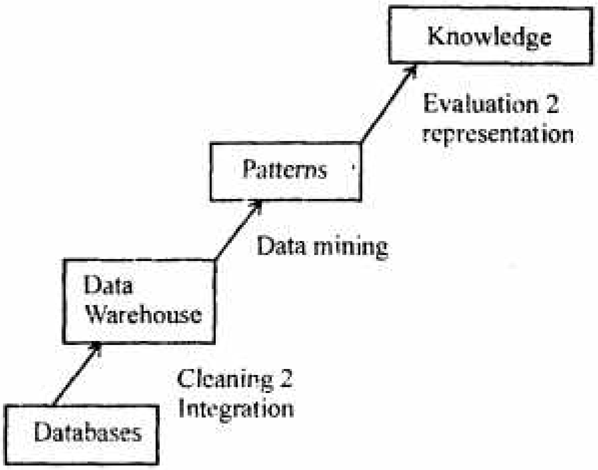
Seventh Semester Examination, May-2009

Data Warehousing & Data Mining (IT-401-E)

Note : Attempt any *FIVE* questions. All questions carry equal marks.

Q. 1. (a) Write the complete process of knowledge discovery from Databases. What is the role played by Data warehousing and Data Mining in this process?

Ans.



Data mining & WH in Process of Knowledge Discover

Steps :

1. Data Cleaning : To remove noise & inconsistent data.

2. Data Integration : Where multiple data sources may be combined.

3. Data Selection : Where data relevant to the analysis task are retrieved from the database.

4. Data Transformation : Where data are transformed or consolidated into forms appropriate for mining by performing summary & aggregation operations.

5. Data Mining : An essential process where intelligent methods are applied in order to extract data patterns.

6. Pattern Evaluation : To identity the truelly interesting patterns representing knowledge based on some interesting measures.

7. Knowledge Representation : Where visualization & knowledge representation techniques are used to present the mined knowledge to the user.

Q. 1. (b) Discuss and differentiate the working of ROLAP, MOLAP and HOLAP servers.

Ans. Relational OLAP (ROLAP) :

- These are the intermediate servers.
- Stand in between a relational back end server & client front end tools.
- Use a relational or extended relational DBMS.
- Servers include optimization for each DBMS back end.

- ROLAP technology tends to have greater stability than MOLAP.

Multidimensional OLAP (MOLAP) :

- Support multidimensional views of data.
- Maps multidimensional views directly to data cube array structures.
- In allows fast indexing to precomputing summarized data.
- Adopt a 2 level storage representation.

Hybrid OLAP (HOLAP) :

- Combine ROLAP & MOLAP.
- Greater stability & scalability of ROLAP.
- Faster computeration of MOLAP.
- Allow large volumes of detail data to be stored in a relational database.
- Microsoft SQL server 7.0 OLAP services support hybrid OLAP server.

Q. 2. (a) Explain the multidimensional data model employed in data warehouse. Highlights its utility.

Ans. Data warehouses & OLAP tools are based as a multidimensional data model. This model views data in form of data cube.

Data Cube : Allows the data to be modelled & viewed in multiple dimensionless.

Example :

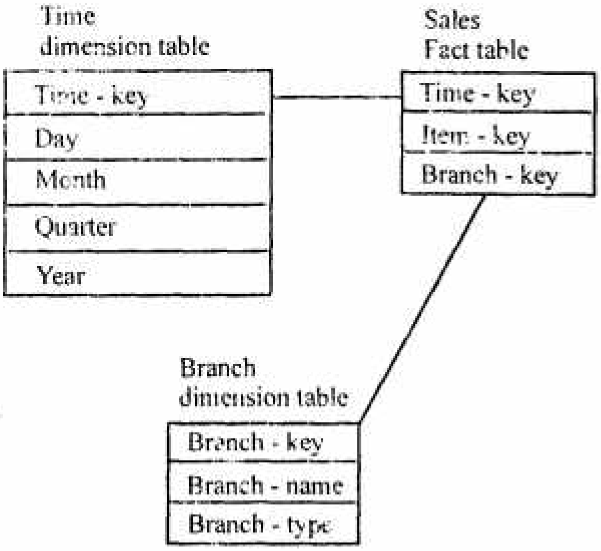


Fig. Star Schema for Multidimensional Data Model

All electronics may create a sales data warehouse in order to keep records of the store's sales with respect to dimensions time, item, branch & location.

A multidimensional data model is typically organized around a central theme, like sales. This theme is represented by a fact table.

Facts are Numerical Measures : The fact table consists of the names of the facts or measures, as well as keys to each of the related dimension tables.

Q. 2. (b) How tuning and testing of data warehouse is carried out, explain.

Ans. Testing & Tuning Data Quality Cross-Fasting Technique :

(i) Customer pains.
(ii) Customer objectives.
(iii) CG solution.
(iv) Cross points.
(v) Customer benefits.

(i) Customer Pains :

- Corporate DWH not being tested in structured way.
- End users only check existing reports with new ones.
- Lack of confidence.
- Delay in getting complete applicant into production.

(ii) Customer Objectives :

- More in depth check on data quality.
- Minimum time effort.
- Easy transfer of testing knowledge.

(iii) City Solution :

- Prove & implement methodology to checks data quality fast & consistent.
- "Dive" in data right away.
- No need to set up complex tex environment.
- No need to gather business knowledge throughout whole company.

(iv) Cross Fooling :

- Comparison between source & target tables.
- Fixed workflow.
- Black box testing.
- Top down approach.

(v) Customer Benefits :

- Increased data quality.
- Faster delivery.
- Fan build up of testing knowledge.

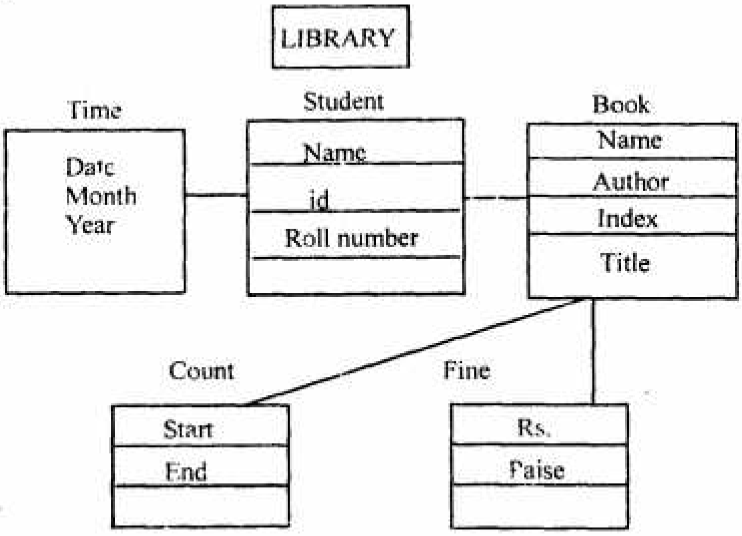
Q. 3. Suppose that warehouse of a big 'LIBRARY' consists of three dimensions time, student and book and two measures count and fine, where count is the number of books issued and fine is the amount (in Rs.) that librarian charges a student for late return of a book.

(i) Draw star and snowflake schema for above data warehouse.

Ans. Conceptual Modeling of Data WH :

Star schemes
Snowflake schemes
Fact conservallions.

Star Schemes :

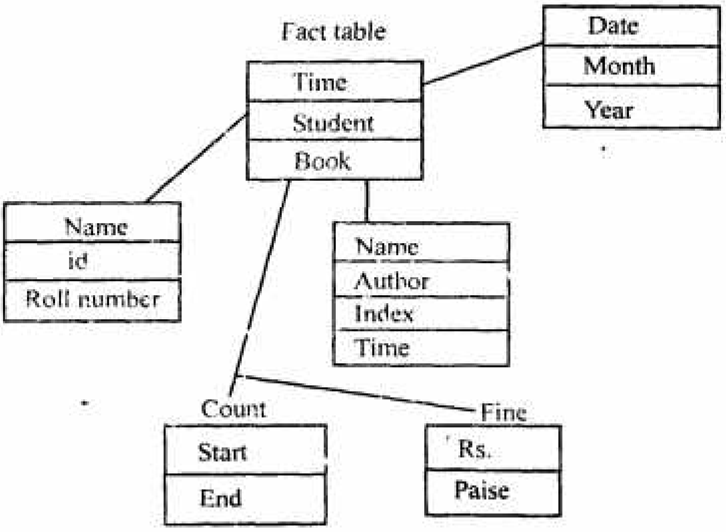


Terms :

- Basic motion.
- Given a collection of numeric measures.
- Each measure depends on a set of dimensions.
- Relation which reacts the dimensionless.
- Each dimensions can have a set of associated attributes.

Q. 3. (ii) Starting with the base cuboid [day, student-name, book-title], what specific OLAP operations should be performed in order to list the total fine paid by each student for books followed in even semesters of 2006.

Ans. Showflake Schemes : A retirement of star schemes where the dimensional hierarchy is represented explicitly by normalizing the dimension table.



Q. 3. (iii) Considering exactly four obspaction levels for each dimension, find the total number of cuboids

Ans. Operation Performed

Step	Activity
1	Closing the process.
2	Choosing the grains.
3	Identify the dimensions.
4	Choosing the facts.
5	Storing the precalculations.
6	Rounding out the dimension tables.
7	Choosing dimension of database.
8	Making slowly changing dimensions.
9	Deciding the query priorities

Of Obspaction Levels : 2 cuboids needed to be constructed.

Q. 4. What is meant by Data Mining Query Langauge? How data specification, knowledge specification, Hierarchy specification and pattern presentation specification can be performed in this language?

Ans. DM query langauge is a desired feature of data mining systems which enables to support adhoc and interactive data mining in order to facilitate flexible & effective knowledge discovery.

Data Specifications :

Use database or use data warehouse. The use clause directs the mining task to the database or data warehouse specified.

From the form & where clauses respectively specification the dB tables or data cubes involved & the continuous defining the data to be retrieved.

Example : Use database all electronics_dB.

```
In relevance to Iname, Iprice, C.income, C.age  from customer C, item I, Purchase P, item_sold S
where    I.item_ID = S.item ID
         and S.transID = P.trans_ID
         and P.curs_ID = C.curs_ID
         and C.Country = "Canadas"
         group by P.date.
```

Knowledge Specification :

Characterization :

```
         <Mine-knowledge_Specification> : : =
         mine characteristics [as <pattern_name>]
         analyse <measure(s)>
mine characteristics as customer purchasing
analysis count %.
```

Discrimination :

<mine_knowledge_specification> : : =

```
mine comparison [as <pattern_name>]
for <target_class> where <target_conditions>
          { versus <contrast_class_i>
  where <contrast_condition_i>}
          analyse <measures)>
```

Hierarchy Specification : Use hierarchy <hierarchy_name> for <attribute_or_dimension>

```
define hierarchy locations_hierarchy on address as
[Streer, city, province-or-state, country].
```

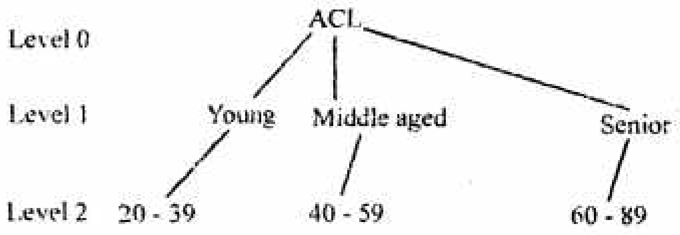


Fig. A concept Hierarchy for Attribue Age

Pattern Presentations :

```
Display as <result-form>
(Multilevel_manipulation> : : = sou up on
                          /drill down on
                          / drop.
```

Q. 5. (a) Describe in detail any one clustering technique.

Ans. Conceptual Clustering :

Typical Requirements :

- Suitability
- Ability to deal with different attributes
- Minimal requirements for domain knowledge
- High dimensionality.
- Constraint based clustering.
- Interpretability.

1. Data Matrix :

$$\begin{bmatrix} x_{i1} & x_{if} & x_{ip} \\ \vdots & \vdots & \vdots \\ x_{i1} & x_{if} & x_{ip} \\ xx_1 & xx_f & xx_p \end{bmatrix}$$

2. Dissimilarity Matrix :

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(x,1) & d(x,2) & .. & 0 \end{bmatrix}$$

Steps :

1. Calculate the mean absolute deviation.

$$s_f = \frac{1}{x}\left[(x_{1f}-m_f)+(x_{2f}-m_f)+\dots+(x_{nf}-m_f)\right]$$

2. Calculate the standarized measurement,

$$Z_{if} = \frac{x_{if}-m_f}{s_f}$$

(a) $d(i, j) \geq 0$

(b) $d(i, i) = 0$

(c) $d(i, j) = d(j, i)$

(d) $d(i, j) \leq d(i, h) + d(h, j)$

$$d(i, j) = \frac{r+s}{q+r+s+t}$$

& $$d(i, j) = \frac{r+s}{q+r+s} \text{ if } t = 0$$

Nominal Variables :

$$d(i, j) = \frac{p-m}{p}$$

m = number of marches.

Q. 5. (b) Explain Fuzzy based techniques of knowledge discovery in data mining.

Ans. Fuzzy Set Approaches : Rule based systems for classification have the disadvantage that they involve sharp cutoffs for continuous attributes.

If (years = employed ≥ 2) $\wedge$ (income $\geq$ 50x) thus,

crldir = "approved".

Such hash thresholding may seen unfair instead fuzzy logic can be introduced into the system to allow "fuzzy" threshold or boundaries to be defined.

Linear & Multiple Regression :

$$Y = \alpha + \beta X$$

$$\beta = \frac{\Sigma_{i=1}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{0}(x_i - \overline{x})^2}$$

$$\alpha = \overline{y} - \beta\overline{x}$$

Multiple regression

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

Non-linear regression,

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

Let $X_1 = X,\ X_2 = X^2 \ \&\ X_3 = X^2$ **Ans.**

Q. 6. (a) Discuss various application areas of data mining in detail.

Ans.

- Graphical user interface.
- Pattern evaluation.
- Data mining engine.
- Data base & DB warehouse server.

Data mining involves an integration of techniques form multiple disciplines such as :

- DB technology.
- Statistics.
- Machine learning.
- High performance computing.
- Pattern recognition.
- Neural networks.
- Data visualization.
- Information retrieval.
- Image & signal processing.
- Spaniel data analysis etc.

Q. 6. (b) What are various back-end tools and utilities in data warehousing, explain?

Ans. 1. Data Extractions : Which typically gathers data from multiple heterogenous sources.

2. Data Cleaning : Which detects errors in the data.

3. Data Transformation : Which converts data from legacy or host format to WH format.

4. Load : Which sorts, summerizes, consolidates, computers views, checks integrity & builds indices & partitions.

5. Refresh : Which propagates the update from the data sources to the warehouse.

Q. 7. (a) Explain the concept of spatial databases. How these databases can be mined, explain?

Ans. Example :

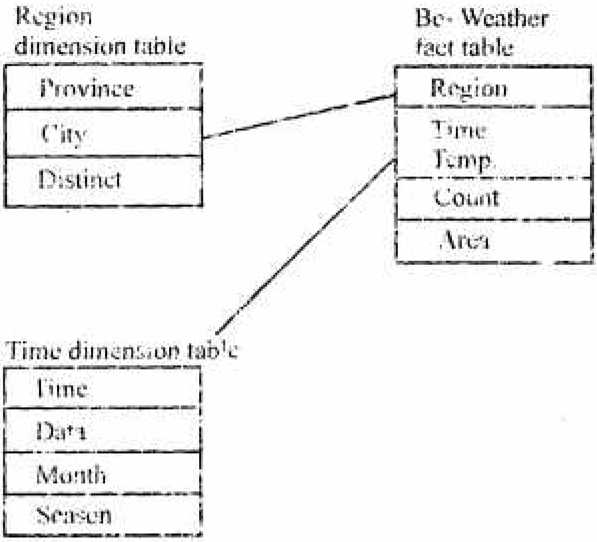


Fig. Spatial DB

Spatial DB contains spatial related information such DB include geographical (map) databases.

Spatial Association Analysis :

Applying certain rough spatial evaluation algorithms. For example, using a minimum bounding rectangle structure and evaluating the relaxed spatial predicate.

Approaches :

- Colour histogram based signature.
- Multifeature composed signature.
- Wavelet based signature.

Q. 7. (b) Describe the mining process of text databases and time-series databases.

Ans. Text DB : Are the DB that contain word descriptions for objects.

Mining Text DB :

Basic measures for text retrieval.

$$\text{Precision} = \frac{|\{\text{Relevant}\} \wedge \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|}$$

$$\text{Recall} = \frac{|\{\text{Relevant}\} \wedge \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|}$$

$$\text{Sim}(V_1, V_2) = \frac{V_1 . V_2}{|V_1||V_2|}$$

Steps :

- Made a term frequency matrix.
- Compute singular valued decomposition.

- Replace the original document vector by a new one.
- Store the set of all vectors 1 craft indices for them using advanced multidimensional indexing techniques.

Time Series DB :

Is an index structure that maintain & hash indexed or B+–tree indexed tables.

- Document-table
- Term-table.

Q. 8. Write short note on any four :

(i) OLAP Query Manager,

(ii) Complex aggregation at multiple granularities,

(iii) Genetic algorithms

(iv) Support vector machines,

(v) Decision trees.

Ans. (i) OLAP Query Manager :

- The purpose of materializing cuboids & constructing OLAP make structures is to speed up query processing in data cubes.

Query processing should proceed as :

1. Determine which operations should be performed on the available cuboids.
2. Determine to which materialized cuboid(s). The relevant operations should be applied.

Example :

Cuboid 1 = Item_name, city year

Cuboid 2 = Brand, country, year

Cuboid 3 = Brand, state, year

Cuboid 4 = Items name. on_state etc

(ii) Complex Aggregation at Multiple Granularities : An important feature of object-relational & melted DB is their capability of storing, accessing & modelling complex structure valued dots such as sen, valued & lisv_valued attribute may be of homogeneous or heterogenous type. Typically set-valued data can be generalised by :

1. Generalising each value in the set into its corresponding higher level concepts.
2. Deviations of the general behaviour of the set.

(iii) Genetic Algorithms : Generic algorithms attempt to incorporate ideas of natural evolution. In general genetic learning starts as follows :

An initial population is created consisting of randomly generated rules. Each rule can be represented by at string of bits.

Based on the nations of survival of the fittest, a new populations is formed to consist of the filters rules in the current population, as well as offspring of these rules.

Off springs are created by applying generic operators such as crossover & nutrition.

In crossover, substrings from pairs of rules are swapped to formed new pairs of rules.

In nutrition, randomly selected bits in a rule's string are inverted.

(iv) Support Vector Machines : The potential usefulness of a pact error is a factor defining its interestingness. it can be estimated by a utility functions, such as supports.

The support of are associated pattern refers to the percentage of task-relevant data tuples for which the patterns is true. For association rules of the form "$A \Rightarrow B$" where A & B are sets of items. It is defined as :

$$\text{Suppose } (A \Rightarrow B) = \frac{\#\text{–tuples cont\_ both A\_ and \_ B}}{\text{Total\_ \# \_ of\_ tuples}}$$

Computer ⇒ Financial_management_software

[Support = 2%, confidence = 60%]

Rule support confidence are 2 measures of rule interestingness that are desired earlier.

(v) Decision Trees : Decision trees were originally intended for classification. Decision tree induction constructs a flow chart like structure where each internal (non-leaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test & each external node denotes a class prediction. At each node, the algorithm chooses the "best" attribute to partitions the data into individuals classes.

If the mining task is classification & the mining algorithms itself is used to determine the attribute subset, then this approach is called wrapper approach, otherwise if in filter approach.