UVA - WELLASSA UNIVERSITY OF SRI LANKA
LIBRARY

UVA WELLASSA UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE & TECHNOLOGY

FIRST SEMESTER EXAMINATION – FEB/MAR – 2012

CST412-3 Data Mining and Data Warehousing

Time allowed: 3 hours

Attach your question paper to the last page of answer script

Answer all questions.

Index No.

Q1.

a. What is meant by data mining? [3 marks]

b. Convert the following relational database into an XML document:

Relation Car_rental

Car_model	Staff_ID	*Trip_ID
MZ-18	A002	T0001
MZ-18	B001	T0002
R-023	B004	T0001
R-023	C001	T0004
SA-38	A001	T0003
SA-38	A02	T0001

Relation Trip

Trip_ID	*Department_ID
T0001	AA001
T0002	AA001
T0003	AB001
T0004	BA001

Relation Department

Department_ID	Salary
AA001	35670
AB001	30010
BA001	22500

[17 marks]

Q2.

a. What are the main differences between operational database systems and data warehouses? [4 marks]

b. 'Surveys indicate that, the use of Data Mining and Data Warehousing technology is becoming increasingly popular in the industry at present.' Explain the reason for this trend. Your answer should include a comparison with traditional database management systems. [8 marks]

c. Discuss, at least two possible areas in Sri Lanka where data mining and data warehousing technology can be applied for the decision making process. A justification of the arguments presented in the answer is required. [8 marks]

Q3.

a. Discuss the importance of Pre-processing of data in data mining and data warehousing industry.

b. Consider the following student marks that are in ascending order:

0,23,32,39,39,40,40,43,45,45,49,52,52,52,60,61,61,69,71,71,71,71,85,92,92,99

i. To smooth the data, the following two data pre-processing methods are to be applied on the above data set.

- Bin means
- Bin boundaries

Assuming number of bins are 4, discuss the **bin mean** and **bin boundaries.**

ii. Using the above data set, briefly describe how one could determine the outliers.

c. Apply Min-max normalization to transform the value 52 for marks onto the range [0.0, 1.0] using the data set given in Part (b) above.

d. Apply Z-score normalization to transform the value 52 for marks, where the standard deviation of marks is 22.64 using the data set given in Part (b) above.

Note:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A.$$

$$v' = \frac{v - mean_A}{stand\_dev_A} \qquad v' = \frac{v}{10^j},$$

[20 marks]

Q4.

a. What are the UML Diagrams available in object oriented design? Define three of them. [6 marks]

b. Given the weather data as shown in the table below:

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

In this table, there are four attributes: outlook, temperature, humidity and wind; and the outcome is whether to play or not.

i. Show the possible Association Rules that can determine the outcome without support and confidence level.

ii. Show the Support level and Confidence level of the following association rule: If temperature = cool then humidity = normal.

[14 marks]

Q5.

a. Compare features between OLAP (On Line Analytical Processing) and OLTP (On Line Transactional Processing) Systems.

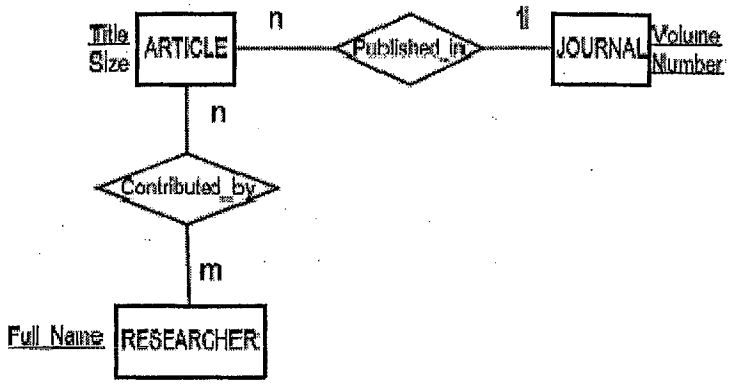
b. Provide an integrated schema for the following two views which are merged to create a bibliographic database. During identification of correspondences between the two views, the users discover the followings:

1. RESEARCHER and AUTHOR are synonyms,
2. CONTRIBUTED_BY and WRITTEN_IN are synonyms,

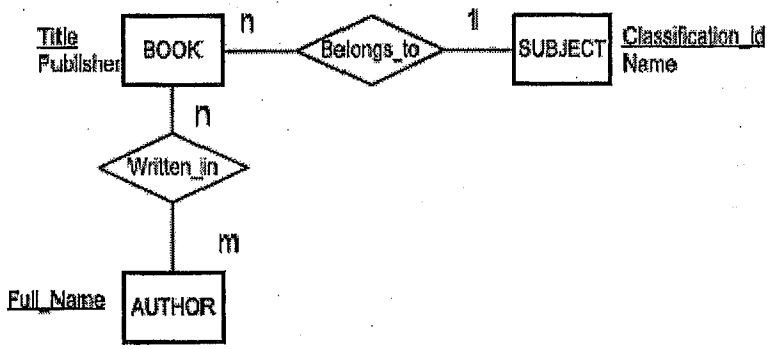
24

3. ARTICLES belongs to a SUBJECT.
4. ARTICLES and BOOK can be generalized as PUBLICATION.

View 1



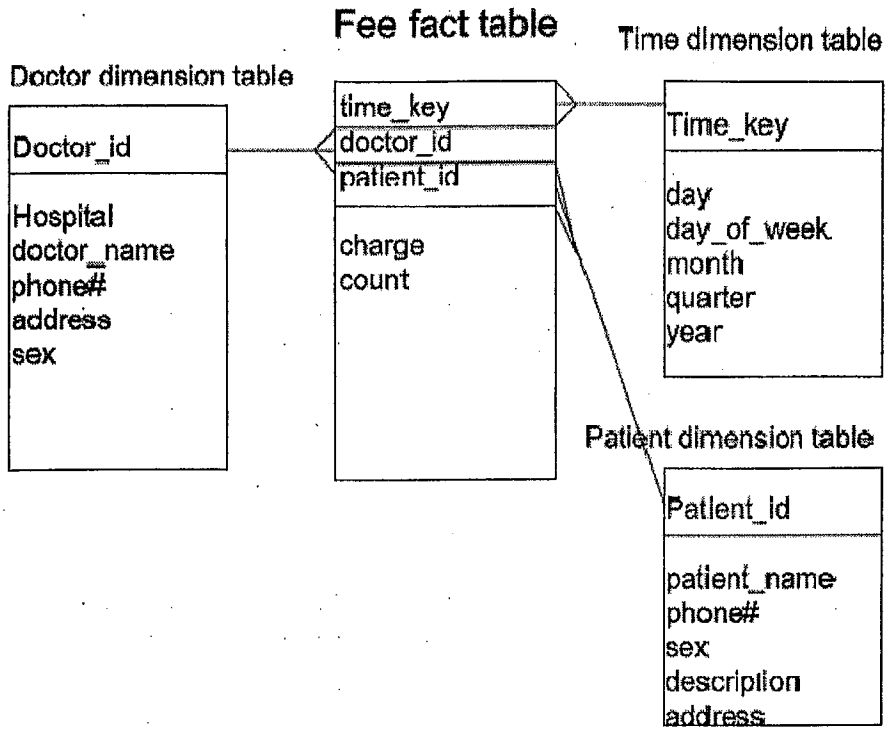
View 2



Hints: Given two subclass entities have same relationship(s). The two subclasses entities can be generalized into a superclass entity and the subclass relationship(s) can also be generalized into a superclass relationship..What is a cube in a data warehouse? Explain how it is constructed and used.

[20 marks]

Q6. Suppose that a data warehouse consists of the three dimensions time, doctor, and patent, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.



a. Starting with the base cuboid [day, doctor, patient], provide a MDX (Multidimensional Expression) query to list the total fee collected by each doctor in 2000?
b. To obtain the same list, write an SQL query assuming the data is stored in a relational database with the table fee (day, month, year, doctor, hospital, patient, count, charge).