UVA WELLASSA UNIVERSITY

FACULTY OF SCIENCE AND TECHNOLOGY

COMPUTER SCIENCE & TECHNOLOGY DEGREE PROGRAM

FIRST SEMESTER EXAMINATION JANUARY/FEBRUARY 2011

CST412-2 DATA WAREHOUSING AND DATA MINING

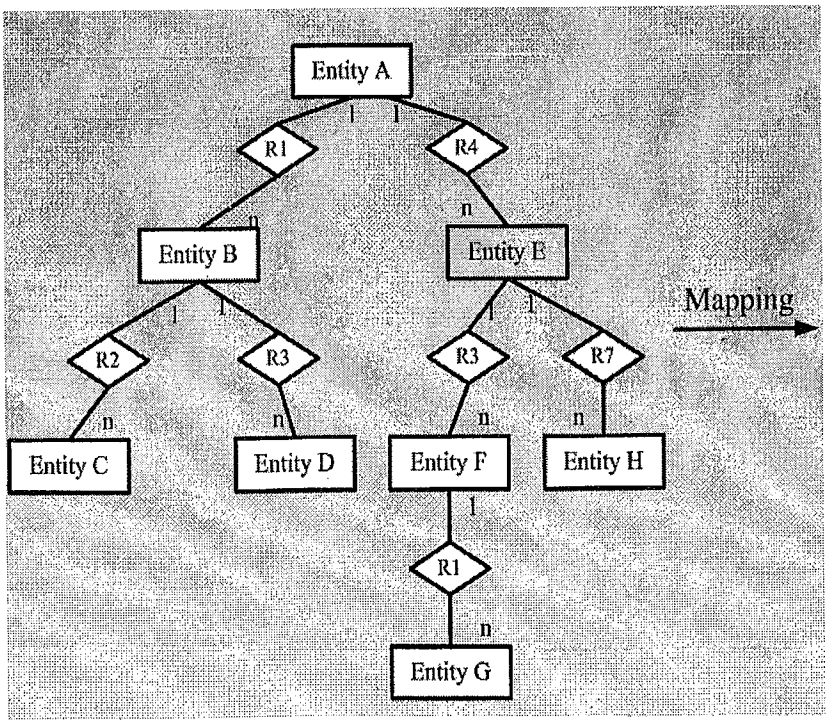
Time allowed: 2 hours	This paper has 6 questions. Answer 5 questions only. You must answer each question on a separate sheet of paper Attach question paper to your answer scripts

Q1.

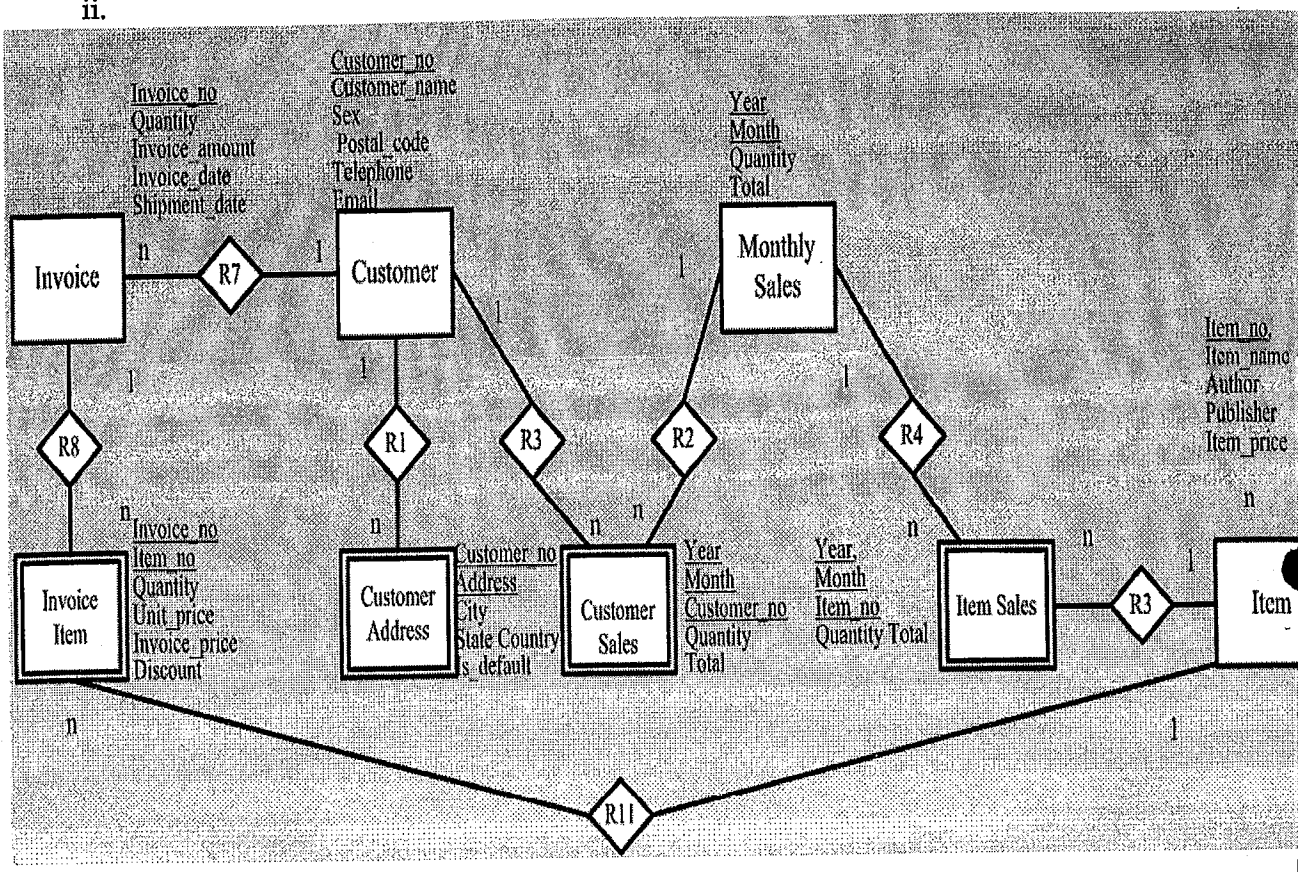
a. Briefly explain about :
   i. XML Syntax
   ii. XML elements
   iii. XML Attributes
   iv. What is a DTD
   v. DTD Graph

b. Mapped these extended entity relationship models to DTD graphs

i.



ii.



c. What are the major differences between generalization and categorization in terms of data volume (data occurrences) in their related superclass entity/entities and subclass entity/entities?

d.

i. Explain stages of reverse engineer relational schema into EER model

ii. In a reverse engineering approach, translate the following relational schema to an Entity-relationship model.

Relation Order	(Order_code, Order_type, Our_reference, Order_date, Approved_date, *Head, *Supplier_code)
Relation Supplier	(Supplier_code, Supplier_name)
Relation Product	(Product_code, Product_description)
Relation Department	(Department_code, Department_name
Relation Head	(Head, *Department_code, Title)
Relation Order_Product	(*Order_code, *Product_code, Qty, Others, Amount)
Relation Note	(*Order_code, Sequence#, Note)

Note: where underlined are primary keys and prefixed with '*' are foreign keys.

20 marks

Q2.

a.

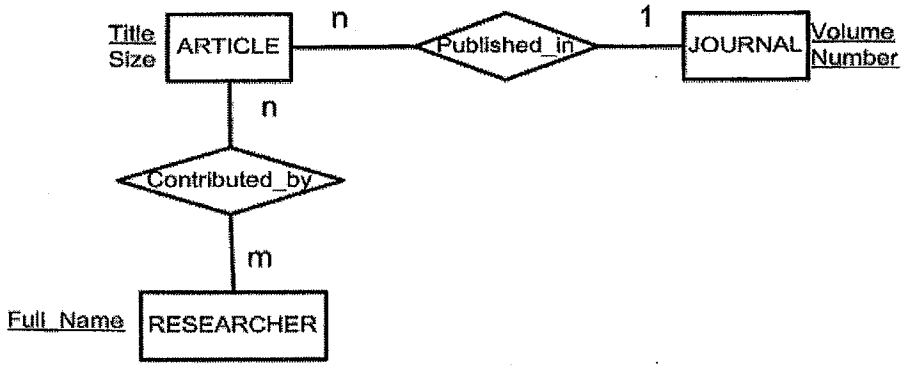
i. Which steps need user supervision for integrating two extended entity relationship models into one extended entity relationship model? Explain your answer.

ii. 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:

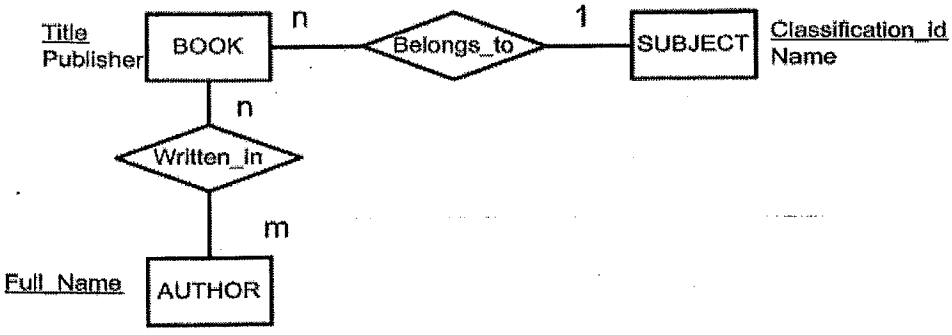
RESEARCHER and AUTHOR are synonyms,
CONTRIBUTED_BY and WRITTEN_IN are synonyms,
ARTICLES belongs to a SUBJECT.
ARTICLES and BOOK can be generalized as PUBLICATION.

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.

View 1



View 2



b.

i. Phases of the decision support life cycle

ii. Compare database with data warehouse in performance, user friendliness, capacity planning and data manipulation language operations?

iii. You are to design a data warehouse to track the sales of salad dressing products in supermarkets at weekly intervals over a four-year period and it is a typical consumer-goods marketing database. The salad dressing product category contains 14000 items at the Universal Product Code (UPC) level. Data are summarized for each of 120 geographic areas (markets) in the United States, and are also summarized for each of 208 weekly time periods spanning over four years. The followings are the tables:

Product Table	(Product_id, Prod_Desc, Brand, Manufacturer, Pack, Class, Flavor, Size)
Sales Table	(*Period_id, *Product_id, *Market_id, Units, Dollars, Discount, Selling_Price, Large_Ads, Medium_Ads, Small_Ads)
Period Table	(Period_id, Period_Desc, Quarter, Fiscal_Year, Calendar_Year, Agg_Level)
Market Table	(Market_id, Market_Desc, District, Region)

Show a simple star schema design for the application.

20 marks

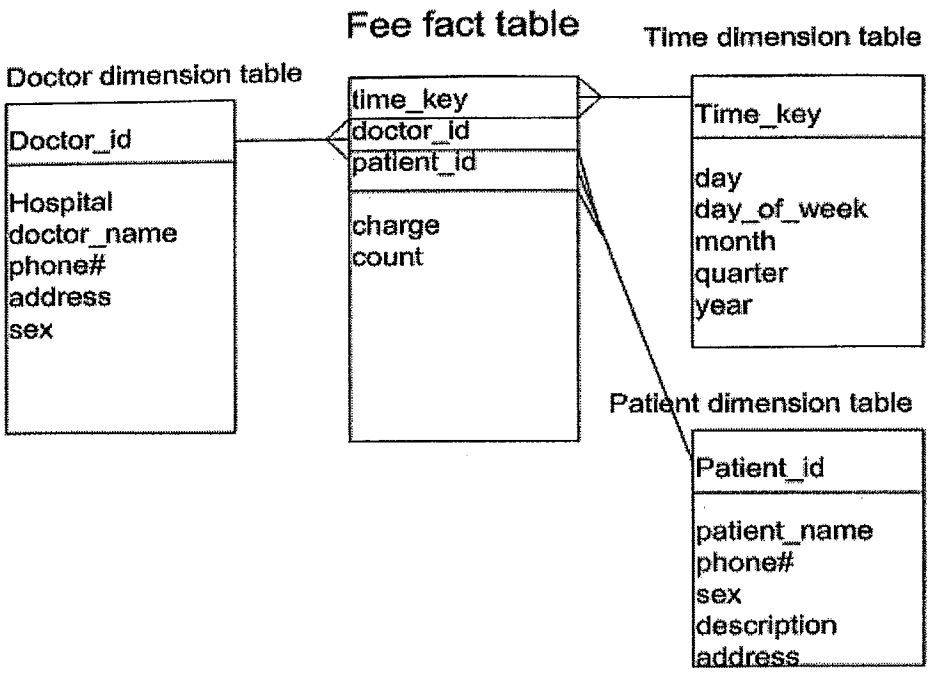
Q3.

a. Explain three methods of implementing an online analytical processing command. Give an example of using one of them with a given Star schema.

b. 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.

i. 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?

ii. 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).*



c. How do you compare the pros and cons of using "Logical Level Translation Approach" with "Customized Program Approach" in data conversion?

d. 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	A002	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

20 marks

Q4.

a. What Is Frequent Pattern Mining?

b. What is the rational of having various data mining technique? In other words, how can one decide which technique of the following to select in data mining?

   i. Association rules
   ii. Clustering
   iii. Decision Tree
   iv. Neural network
   v. Web Mining
   vi. Genetic programming

c. What are the major differences between Apriori algorithm and Frequent Pattern Tree (FP-tree) with respect to performance? Justify your answer.

d. 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.

20 marks

Q5.

a.

i. Define Forward maximal sequence, with its algorithm and with is its application on customer relationship management in e-commerce.

ii. Find the maximal forward references of web pages in a database D of sessions (A, B, C), (A, C, B), (B, C, E), (A, C), (A, C, D, C, E) and (A, B, C, A, C, B, C, A, C, D, E) with the minimum support S_{min} of two sessions

iii. What is the termination condition of growing tree phase?

b.

i. What is Cluster analysis? give three examples of Clustering applications

ii. Explain The *K-Means* clustering method with example

iii. What is supervised clustering and what is unsupervised clustering? How do you compare their difference with respect to performance? Illustrate the strength and weakness of k-means in comparison with the k-medoids algorithm

iv. The following table contains the attributes name, gender, trait -1(characteristic1), trait-2 (characteristic2), trait-3(characteristic3), and trait-4(characteristic4),, where name is an object-id, gender is a symmetric attribute, and the remaining trait attributes are asymmetric, describing personal traits of individuals who desire a penpal. Suppose that a service exists that attempts to find pairs of compatible penpals. For asymmetric attribute values, let the value P be set to 1 and the value N be set to 0. Suppose that the distance between objects (potential penals) is computed based only on the asymmetric variables.

01) Compute the Jaccard coefficient for each pair.

02) Who do you suggest would make the best pair of penpals? Which pair of individuals would be the least compatible?

Name	Gender	Trait-1	Trait-2	Trait-3	Trait-4
Kevan	M	N	P	P	N
Caroline	F	N	P	P	P
Erik	M	P	N	N	P

20 marks

Q6. Given the following training data set

Training instance	Income range	Credit card insurance	Sex	Age
1	30-40k	Yes	Male	30-39
2	30-40k	No	Female	40-49
3	50-60k	No	Female	30-39
4	20-30k	No	Female	50-59
5	20-30k	No	Male	20-29
6	30-40k	No	Male	40-49

a. Describe the steps needed to apply unsupervised genetic learning to cluster the instances of the credit card promotion database.

b. After transforming the input data into numeric such as yes=1, no=2, male=1, female=2, 20-29=1, 30-39=2, 40-49=3, 50-59=4, 20-30k=1, 30-40k=2, 40-50k=3, 50-60k=4, the training data set becomes:

T(1)=(2,1,1,2)
T(2)=(2,2,2,3)
T(3)=(4,2,2,2)
T(4)=(1,2,2,4)
T(5)=(1,2,1,1)
T(6)=(2,2,1,3)
Assume there are two set of initial population for two clusters as:
Solution 1 of 2 clusters centers: K1(1,1,1,1), (4,2,2,4)
Solution 2 of 2 clusters centers: K2(4,4,4,4), (2,2,1,1)

By using above training data set choose the best solution based on their fitness function score by use of unsupervised genetic learning.

20 marks