THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF COMPUTING

EXAMINATION

Course : MSc Scheme - 61030

Subject : COMP5121 Data Mining And Data Warehousing Applications

Group : 101, 102, 103, 104, 105

Session : 2010 / 2011 Semester I

Date : 10 December 2010 Time : 18:30-20:30

Time Allowed: 2 Hours Subject Lecturer: Korris Chung

This question paper has ___4___ pages (cover included).

Instructions to Candidates:

Open book examination.
Answer ALL questions.
Only standard calculator is allowed.

Do not turn this page until you are told to do so!

THE HONG KONG POLTECHNIC UNIVERSITY

COMP5121 Data Mining & Data Warehousing Applications

2010-11 Fall Term Examination

Instructions:
Open book examination.
Answer ALL FIVE questions and show your steps.
Write down the assumption(s) you made when necessary and interpret the questions logically.
Only standard calculator is allowed.

1. Given the following survey data with missing values as a result of privacy concerns which include the followings. Elaine wants to hide her title. Grace wants to hide her educational background. Alvin wants to hide his marriage status and Katherine wants to hide her age information as well as her salary level. You are asked to make use of the association rule mining technology to guess the missing data values.

Table I Survey Data with missing values as a result of privacy concerns

Name	Age	M. Status	Education	Title	Salary Level
Andy	30-40	Unmarried	HD	Manager	SL-3
Brian	20-30	Married	HD	Assistant	SL-3
Donald	20-30	Unmarried	University	Manager	SL-5
Elaine	40-60	Unmarried	University		SL-7
Franky	40-60	Married	University	Manager	SL-5
Grace	30-40	Married		Manager	SL-7
Helsa	30-40	Married	M.B.A.	Manager	SL-5
Alvin	20-30		M.B.A.	Accountant	SL-5
Yvonne	30-40	Unmarried	University	Accountant	SL-5
Katherine		Married	University	Assistant	

a) Describe how you can make use of association rule mining to mine on the data in Table I so that the missing data values can be properly guessed. You may need to list the database records and attributes to be used by association rule mining algorithms like Apriori. You may also need to show how to use the association rules mined in guessing the missing values.

(6 marks)

b) Based on your idea in part (a), show how the title of Elaine is predicted. Note here that you are NOT required to show all the steps in the association rule mining process. Your answer may just include the followings
   - the association rules that are relevant to the prediction of Elaine's title
   - the prediction of Elaine's title based on the relevant association rules

   Justify your answer.

(14 marks)

2. Association rule mining is not the only method to guessing the missing data values. Suppose you are further asked to make use of classification technologies to predict the missing values in Table I.

   a) Describe how you can solve the problem. You may need to list the database records and attributes to be used by the adopted classifier.

   (6 marks)

   b) Apply your idea in part (a) and use the naive Bayesian classifier to predict the age of Katherine. Show also the conditional probability values of the Bayesian classifier you construct.

   (14 marks)

3. Given the following tourist data for sequential association rule mining. Each row records the city/cities visited in a 3-month period by a tourist.

Table II Tourist database

Tourist ID	Time Period	City/Cities Visited
T0001	Jan-Mar 2006	Phuket, Tokyo
T0001	Jan-Mar 2007	L.A.
T0001	July-Sept 2008	Seoul
T0002	Jan-Mar 2008	Phuket,Tokyo,Seoul
T0003	Jan-Mar 2006	Phuket
T0003	July-Sept 2008	Tokyo
T0003	Oct-Dec 2008	Seoul
T0004	Oct-Dec 2005	Sydney, Phuket
T0004	Oct-Dec 2006	Tokyo
T0004	July-Sept 2007	L.A.,Hokkaido
T0004	Jan-Mar 2009	Tokyo,Vancouver

   a) What is the maximum length of the frequent sequences (i.e. the maximum number of itemsets in a frequent sequence) for minimum support *min_sup*>0%? Note here that you don't need to apply any sequential association rule mining algorithm.

   (3 marks)

   b) What is the largest size of the frequent itemsets (i.e. the maximum number of items in a frequent itemset) found by frequent itemset phase of the sequential association rule mining for minimum support *min_sup*>0%? Note here that you don't need to apply any itemset mining here.

   (3 marks)

   c) Find all frequent sequences of length less than or equal to 2 using sequential association rule mining (AprioriAll algorithm) for *min_sup*=40%. Show your mining steps, including the transformation phase and C_1, L_1, C_2, & L_2 in the sequence phase.

   (14 marks)

4. You are asked to further analyze the tourist database in Table II of Q.3 by grouping the tourists according to their similarity/dissimilarity.

   a) Propose an appropriate dissimilarity measure for this database and prepare the following dissimilarity matrix for clustering.

	*T*0001	*T*0002	*T*0003	*T*0004
*T*0001	0	–	–	–
*T*0002	_____	0	–	–
*T*0003	_____	_____	0	–
*T*0004	_____	_____	_____	0

(10 marks)

   a) Based on the completed dissimilarity matrix in part (a), cluster the data records using the single linkage agglomerative hierarchical clustering algorithm. Draw the dendrogram found.

(10 marks)

5. Suppose the electronic health record system initiated by the HKSAR government includes a health data warehouse and you are assigned to design it. Consider the scenario that there are four dimensions: (i) patient, (ii) disease, (iii) hospital/clinic/lab, and (iv) time, and two measures: *dollar_cost* and *number_of_case* where *dollar_cost* is the cost involved and *number_of_case* is the number of such cases.

   a) Design a star schema for the above data warehouse. You may design your own dimension attribute names.

(6 marks)

   b) Discuss the usage of your designed data warehouse by listing 4 questions that can be answered.

(4 marks)

   c) Suppose you start from the highest level of summarization, i.e. 0-D (apex) cuboid. What OLAP operations will be needed to find the number of cases of a particular disease in a particular hospital?

(5 marks)

   d) Briefly discuss the difference between roll-up and slicing OLAPs. Will there be a case that they will come up with the same result? Justify your answer.

(5 marks)

—END—