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, 1888

Session : 2008 / 2009 Semester I

Date : 15 December 2008 Time : 18:30-20:30

Time Allowed: 2 Hours Subject Lecturer: Korris Chung

This question paper has ___6___ pages (cover included).

Instructions to Candidates:

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.

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

COMP5121 Data Mining & Data Warehousing Applications

2008-09 Final Examination

1. Given the following modified weblog data for sequential association rule mining. Each row records the URLs (web addresses) visited within a particular time frame by a visitor. For example,

Visitor	Time Frame	URLs Requested within Time Frame
IP-00001	20:00-20:15, 2 Dec 2008	30,50
IP-00001	21:30-21:45, 2 Dec 2008	40
IP-00001	20:00-20:15, 3 Dec 2008	70
IP-00002	20:00-20:15, 2 Dec 2008	10,30
IP-00002	20:00-20:15, 12 Dec 2008	50
IP-00002	20:30-20:45, 12 Dec 2008	40,60
IP-00002	22:00-22:15, 13 Dec 2008	50,90
IP-00003	20:00-20:15, 2 Dec 2008	30,50,70
IP-00004	23:00-23:15, 5 Dec 2008	30
IP-00004	23:15-23:30, 5 Dec 2008	50
IP-00004	23:30-23:45, 5 Dec 2008	70

a) List all possible sequences arising from the visitor IP-00001. You may assume that events can be repeated in a sequence.

(3 marks)

b) 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 here.

(2 marks)

c) 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.

(2 marks)

d) Find all frequent sequences of length less than or equal to 3 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, C_3, & L_3 in the sequence phase.

(13 marks)

2. Your R&D team has been assigned a project to carry out opinion mining on the on-line forum postings. After pre-processing the collected text documents, the following sample database is given, with the overall opinion column considered as the class attribute.

Database for Opinion Mining

Document ID	**A1**:Number of Sentences with Positive Opinion	**A2**:Number of Sentences with Negative Opinion	**A3**:Happy Smile Code Used	Overall Opinion
F1-D1	*Less*	*Less*	*No*	*Negative*
F1-D2	*Less*	*Intermediate*	*No*	*Neutral*
F1-D3	*Intermediate*	*More*	*Yes*	*Positive*
F2-D1	*More*	*Intermediate*	*Yes*	*Positive*
F2-D2	*Intermediate*	*Less*	*Yes*	*Negative*
F2-D3	*Less*	*Less*	*No*	*Neutral*
F2-D4	*More*	*More*	*Yes*	*Positive*

a) Suppose you are asked to adopt the naive Bayesian classifier to classify the overall opinion. Compute the classification rate of the training dataset if the seven documents above are used for training.

(10 marks)

b) You are further asked to use the following association rules mined from the database above to predict the overall opinion of unseen documents.

A1=More → Overall Opinion=Positive [s=2/7]
A2=More → Overall Opinion=Positive [s=2/7]
A3=Yes → Overall Opinion=Positive [s=3/7]
A1=More, A3=Yes → Overall Opinion=Positive [s=2/7]
A2=More, A3=Yes → Overall Opinion=Positive [s=2/7]
A1=Less → Overall Opinion=Neutral [s=2/7]
A3=No → Overall Opinion= Neutral [s=2/7]
A1=Less, A3=No → Overall Opinion= Neutral [s=2/7]
A2=Less → Overall Opinion=Negative [s=2/7]

Describe how you solve this problem and show how the following two records are classified.

Document ID	Number of Sentences with Positive Opinion	Number of Sentences with Negative Opinion	Happy Smile Code Used	Overall Opinion
F1-D4	*More*	*Intermediate*	*Yes*	?
F2-D5	*Less*	*Intermediate*	*Yes*	?

(10 marks)

3. Given the following tourist data records, which show the temporal sequences of cities visited in the past five years. You are assigned a data mining task for this data set, i.e., to cluster the tourists/visitors based on the given travel patterns.

Tourist ID	Sequence of Cities visited in the past five years
20001	Bangkok → Tokyo → Osaka → Beijing → London
20002	Osaka → Tokyo → Beijing → Osaka
20003	Beijing → Osaka → Tokyo → Bangkok → Toronto
20004	Beijing → Taipei → Osaka → Bangkok → Toronto

a) Propose a dissimilarity measure for this clustering task by taking into consideration the provided sequential information. Compute and fill in the missing values of the dissimilarity matrix below.

	20001	20002	20003	20004
20001	0			
20002	_____	0		
20003	_____	_____	0	
20004	_____	_____	_____	0

(14 marks)

b) Based on the completed dissimilarity matrix in part (a), show how the data records are grouped by the single linkage agglomerative hierarchical clustering algorithm. You may JUST show the first merging of two tourists.

(6 marks)

4. Suppose you are asked to provide data mining consulting services to a restaurant review website. After interviewing the company's manager and the database administrator, the following sample database recording the member's review of restaurants are collected.

Restaurant Review Database

Member ID	Review Date	Restaurant Reviewed	Rating (5-star scheme)		
			Food	Service	Environment
00001	02-01-2003	Steak Xpert	★★★	★★	★★★
00001	12-11-2003	Deleefrance	★★	★★★	★★★
00001	15-05-2006	Maxiim	★★	★★★	★★
00002	12-01-2003	Hung Sing	★★★★	★★★	★★★
00002	12-10-2004	Maxiim	★★★	★	★
00002	10-06-2005	Sushi King	★★★★	★★★★	★★★
00003	07-03-2005	MaCafe	★★★	★★	★★★★
00003	16-03-2006	Cafe DeCora	★★	★	★
00004	07-03-2005	Super Dumbling	★★★	★	★
00004	17-03-2006	Curry in a Box	★	★★	★
00004	18-03-2006	MaCafe	★★	★	★★★

a) If you are asked to provide restaurant recommendation service to the company's members, describe how you formulate and solve the problem. Note that you are free to use any data mining technology to accomplish this task. You may also recommend collecting other information to improve the effectiveness of your solution.

(10 marks)

b) If you are further asked to form member group of similar restaurant preference for appropriate online restaurant discussion forums, describe how you formulate and solve the problem. Again, you are free to use any data mining technology to accomplish this task and are allowed to recommend collecting other information to improve the effectiveness of your solution.

(10 marks)

5. a) Suggest an effective method to determine the missing values below. Fill in the missing values accordingly.

Patient ID	Blood Pressure Level	Sex	Age	Fever	Disease
9100123	80-120	Male	65	Yes	No
9303034	160-200	Female	55	No	Yes
9210126	80-120	Male	12	Yes	Yes
9142020	120-160	Female	35	No	No
9910111	160-200	Male	46	No	Yes
9576732	80-120	Male	16	Yes	No
9910115	160-200	Female		No	Yes
9210120	120-160		23		
9576737			28	Yes	No

(6 marks)

b) Suppose you are responsible to design a data warehouse for hospital authority (HA) and are given the following five dimension tables and two fact table measures: charge and cost where charge is the amount paid by patient and cost is the amount spent by HA.

Dimension Tables:

- **Time**
  for queries comparing charge/cost by season, quarter, or holiday
- **Medical Service**
  for queries comparing charge/cost by type of medical services
- **Hospital**
  for queries comparing charge/cost by hospital or hospital network
- **Patient**
  for queries comparing charge/cost by different age group, sex, or family background
- **Doctor**
  for queries comparing charge/cost by different specialty, experience, or rank

i) Design a galaxy schema for this data warehouse. You may design your own dimension attribute names.

(8 marks)

ii) Suggest two queries that can be applied to this data warehouse.

(6 marks)

—END—