THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF COMPUTING

EXAMINATION

Course : MScIS (60130/61020) / MScEC (61030/61027/61028) / Research Student

Subject : COMP5121 Data Mining and Data Warehousing Applications

Group : 103,1031 / 104,1041 / 1888

Session : 2005 / 2006 Semester I

Date : 15 December 2005 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 any **FOUR** questions. Each question carries equal marks.
Show all your steps and write down any assumption(s) you made.

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

Answer ANY FOUR questions. Each question carries 25 marks.

1. Given the following modified anonymous weblog data from msnbc.com (http://kdd.ics.uci.edu/databases/msnbc/msnbc.html) for sequential pattern mining.

```
% Sequences:
1 1
2
3 2 2 4 2 2 2 3 3
5 6 7 5 6 7
1 11 6 11 1
6 8 7 8 7
1 1 8 7 6
6 6 7 8
6 7 7 7 6 6 8 8 8 8
6 9 4 4 4 10 3 10 5 10 3 4 4 4
1 1 1 11 1 1 1
12 12
1 1 3 3
```

where each sequence (line) corresponds to page views of a user during a particular time period. Each event in a sequence corresponds to a user's request for a page recorded at the level of page category. The categories are labeled as:
1: "frontpage", 2: "news", 3: "tech", 4: "local", 5: "opinion", 6: "on-air", 7: "misc", 8: "weather", 9: "msn-news", 10: "health", 11: "living", 12: "business", 13: "msn-sports", 14: "sports", 15: "summary", 16: "bbs", 17: "travel".

The interpretation of this sequence file is straightforward as exemplified below.

- The first user, corresponding to the first line under % Sequence, hits "frontpage" and then "frontpage", i.e., 1 1.
- The third user, corresponding to the third line under % Sequence, hits "tech" (3), then "news" (2), "news" 2, then "local" (4), and then "news" for three consecutive times (2 2 2), and finally "tech" for two more times (3 3).

a) List all possible sequences arising from the sixth user, i.e. 6 8 7 8 7. You may assume that events can be repeated in a sequence.

(5 marks)

b) For the two sequences <(6),(7),(8)> and <(4),(3),(3)>,
   i) compute their supports.
   ii) generate all strong sequential association rules using minimum confidence=50%. Assume that these two sequences are frequent.

(10 marks)

c) What is the maximum length (i.e. the maximum number of itemsets) of the frequent sequences for minimum support $min\_sup \geq 0\%$? Note here that you don't need to apply the AprioriAll algorithm.

(2 marks)

d) What is the maximum size of the frequent itemsets in the frequent itemset phase (i.e. Step 2) of the sequential pattern mining process? Again, $min\_sup \geq 0\%$.

(2 marks)

e) Suggest three potential uses of the sequential association rules mined from this weblog database for msnbc.com. Elaborate your answer.

(6 marks)

2. Given the following university's undergraduate student database records.

Name	A1 Gender	A2 Graduated from band 1 High School	A3 Active Participation in Extra-curriculum Activities	A4 Got 1st Class Honor
Alpha	M	Y	N	N
Bethia	F	Y	N	Y
Carol	F	Y	Y	N
Daisy	F	N	N	N
Eten	M	Y	N	Y
Finn	M	N	Y	Y

a) Suggest a distance metric for clustering the records above. Compute and fill in the missing values of the distance matrix below.

	Alpha	*Bethia*	*Carol*	*Daisy*	*Eten*	*Finn*
Alpha	0					
Bethia	___	0				
Carol	___	___	0			
Daisy	___	___	___	0		
Eten	___	___	___	___	0	
Finn	___	___	___	___	___	0

(7 marks)

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

(9 marks)

c) With respect to your answer in part (b), how many clusters can be formed?

(4 marks)

d) Suppose now some of the attributes are changed to categorical type as follows, i.e. A1 and A3 are binary while A2 and A4 are categorical. Compute the distance between (i) Alpha and Bethia, and (ii) Daisy and Eten.

Name	A1 Gender	A2 Band of Sec. School Graduated from	A3 Participation in Extra-curriculum Activities	A4 Class of Honor Awarded
Alpha	M	A	N	I
Bethia	F	B	N	II
Carol	F	A	Y	III
Daisy	F	C	N	I
Eten	M	C	N	II
Finn	M	B	Y	II

(5 marks)

3. Consider the following stock price movement data.

Stock	Price Movement from 12 December – 23 December, 2005									
	12/12	13/12	14/12	15/12	16/12	19/12	20/12	21/12	22/12	23/12
PCCW	*Up*	*Up*	*Level*	*Down*	*Level*	*Up*	*Up*	*Down*	*Level*	*Up*
HSBC	*Down*	*Down*	*Down*	*Up*	*Level*	*Level*	*Down*	*Up*	*Up*	*Down*
CTI	*Level*	*Level*	*Up*	*Up*	*Level*	*Down*	*Level*	*Level*	*Level*	*Up*

where the movement labels *Up*, *Down* & *Level* denote the stock price going up, down and level respectively in the corresponding trading day, PCCW and CTI are two telecommunication stocks, and HSBC is a banking stock.

a) In order to classify next trading day's price movement for PCCW, the following data are extracted.

Today is	Price Movement of HSBC for			
	2 Trading Day before	1 Trading Day before	Today	Next Trading Day
14 Dec.	*Up*	*Up*	*Level*	*Down*
15 Dec.	*Up*	*Level*	*Down*	*Level*
16 Dec.	*Level*	*Down*	*Level*	*Up*
19 Dec.	*Down*	*Level*	*Up*	*Up*
20 Dec.	*Level*	*Up*	*Up*	*Down*
21 Dec.	*Up*	*Up*	*Down*	*Level*
22 Dec.	*Up*	*Down*	*Level*	*Up*

Suppose you take use of the naive Bayesian classification method to solve the problem and all the seven records above are used for training.

i) Compute the classification rate of the training data.

(12 marks)

ii) Show how the following data should be classified.

Today is	2 Trading Day before	1 Trading Day before	Today
23 Dec	*Down*	*Level*	*Up*

(3 marks)

b) Comment on whether the CTI and/or HSBC price movement data should be added to part (a) to obtain more data and consequently more accurate classification.

(3 marks)

c) Suggest another way to classify the stock price movement of PCCW or HSBC or CTI above.

(7 marks)

4. As a data mining consultant, you are working for a DVD shop to design a system to provide movie recommendation based on the movie ratings expressed by the customers, as exemplified below.

Customer	Movie Information	Rating (5-stars scheme)
A	Beautiful Mind, A (2001) Drama, Mystery, Romance	4.5 stars
	Poltergeist (1982) Horror, Thriller	3 stars
	Batman Begins (2005) Action, Crime, Thriller	2 stars
	Fly, The (1986) Drama, Horror, Sci-Fi	2.5 stars
	Ghostbusters II (1989) Action, Comedy, Fantasy, Horror, Sci-Fi	4 stars

You may assume that a 5-stars rating scheme is adopted and the movie information is limited to the year of production and the movie categories like drama, romance, action, thriller, comedy, etc.

a) If the classification technology is adopted, describe how you formulate and solve the problem. You are expected to limit your discussions to the descriptions above and answer the following questions.
   i) What is the class attribute? What are the ordinary attributes for classification?
   ii) What classification model(s) will you suggest?
   iii) How can the classification knowledge be used?

(9 marks)

b) If the clustering technology is adopted, describe how you formulate and solve the problem. Again, you are expected to limit your discussions to the descriptions above and answer the following questions.
   i) What database attributes will you use for clustering?
   ii) What clustering technique(s) will you suggest?
   iii) How can the clustering knowledge be used?

(8 marks)

c) If the association rule mining is adopted, describe how you formulate and solve the problem. You are expected to limit your discussions to the descriptions above and answer the following questions.
   i) How should a transaction and an item be defined? Give a few examples.
   ii) What association rules do you expect?
   iii) How can the association rules be used?

(8 marks)

5. a) Suggest a binning method to filter out the noisy data below with 3 bins.

-100, 3, 4, 5, 50, 90, 95, 100, 120, 125, 130, 1000

(4 marks)

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

Customer ID	Monthly Income	Age	Education	Marital Status	Usage
9100123	Low	Old	University	Married	Low
9303034	High	Young	College	Single	High
9210126	Medium	Young	College	Married	High
9142020	Medium	Old	High School	Single	Low
9910111	High	Old	University	Single	High
9576732	Low	Old	High School	Married	Low
9910115		Young	University	Single	High
9210120	Medium		College		Low
9576737	Low	Young		Married	Low

(5 marks)

c) Suppose you are responsible for designing a data warehouse for the hospital authority (HA) and are given three dimensions: (i) disease, (ii) hospital/clinic, and (iii) time, and two measures: *dollar_cost* and *number_of_case* where *dollar_cost* is the cost of handling a disease in a hospital/clinic during a period of time and *number_of_case* is the number of such disease case in a hospital/clinic during a period of time.

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

(7 marks)

ii) List 3 questions/hypotheses that can be answered/confirmed by querying your design in part (c-i). Write down the necessary OLAP steps. Make your own assumption(s).

(9 marks)

- E N D -