THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF COMPUTING

EXAMINATION

Course : MScIT (61030/88004) / MScST (61030/88004) / Research Student

Subject : COMP5121 Data Mining And Data Warehousing Applications

Group : 201,2011/ 202,2021/ 205/2888

Session : 2005 / 2006 Semester II

Date : 10 May 2006 Time : 18:30-20:30

Time Allowed : 2 Hours Subject Lecturer : Korris Chung

This question paper has ___7___ 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. Suppose you are asked to provide data mining consulting services to an Internet DVD shop. After interviewing the shop's manager and the database administrator, the following information about the customer database and the movie database are collected.

Customer Database

Customer ID	Transaction Date	Movie Rent (Movie ID)	Rating (5-star scheme)
00001	02-01-2003	3997 (Spiderman II)	4 stars
00001	12-11-2003	0553 (Finding Nemo)	3.5 stars
00001	15-05-2006	0150 (Cinderella Man)	4 stars
00002	12-01-2003	1011 (Poltergeist)	4 stars
00002	12-10-2004	0150 (Cinderella Man)	3 stars
00002	10-06-2005	3996 (Spiderman)	3.5 stars
00003	07-03-2005	0013 (Batman Begins)	3.5 stars
00003	16-03-2006	0001 (A Beautiful Mind)	2 stars
00004	07-03-2005	4490 (The Fly)	3.5 stars
00004	17-03-2006	0909 (King Kong)	5 stars
00004	18-03-2006	0013 (Batman Begins)	4 stars

Movie Database

Movie ID	Movie Name	Types
0001	A Beautiful Mind	Drama, Mystery, Romance
0012	Batman	Action, Crime, Thriller
0013	Batman Begins	Action, Crime, Thriller
0150	Cinderella Man	Drama, Romance
0553	Finding Nemo	Animation, Comedy
1011	Poltergeist	Horror, Thriller
3996	Spiderman	Action, Crime, Sci-Fiction
3997	Spiderman II	Action, Crime, Sci-Fiction
4490	The Fly	Drama, Horror, Sci-Fiction
0909	King Kong	Action, Thriller, Horror, Sci-Fiction

If the clustering technology is adopted, describe how you formulate and solve the problem as follows.

a) Compute the similarities between the ~~five~~ four customers so that they can be effectively clustered according to their movie preferences. You are expected to take into considerations of the customer's rating and movie types in your formulation.

(10 marks)

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

(7 marks)

c) Your client would like to know the movie preferences of the atypical customers. Discuss how the dendrogram found by the agglomerative hierarchical clustering algorithm can help to identify such kind of customers, i.e. the outliers.

(5 marks)

d) Knowing that the agglomerative hierarchical clustering is computationally expensive, propose a method to speed up the algorithm.

(3 marks)

2. Consider the following stock price movement data:

Stock Price Database

Stock	Price Movement from 13 March – 24 March, 2006									
	13/3	14/3	15/3	16/3	17/3	20/3	21/3	22/3	23/3	24/3
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.

a) Carry out an association analysis of the data below extracted from the stock price database above.

Extracted Stock Price Database for Association Analysis

Today is	Price Movement of PCCW for			
	2 Trading Day before	1 Trading Day before	Today	Next Trading Day
15 Mar.	*Up*	*Up*	*Level*	*Down*
16 Mar.	*Up*	*Level*	*Down*	*Level*
17 Mar.	*Level*	*Down*	*Level*	*Up*
20 Mar.	*Down*	*Level*	*Up*	*Up*
21 Mar.	*Level*	*Up*	*Up*	*Down*
22 Mar.	*Up*	*Up*	*Down*	*Level*
23 Mar.	*Up*	*Down*	*Level*	*Up*

By setting the minimum support to 25% and the minimum confidence to 50%, mine all strong association rules satisfying

* → Next Trading Day=Up

or * → Next Trading Day=Down

or * → Next Trading Day=Level

where * denotes a wild card. Note here that you are NOT required to apply the Apriori algorithm to generate the solution (otherwise you will use up your time).

(10 marks)

b) You are further asked to use the rules mined in (a) to predict next trading day's price movement.

i) Describe how you solve this problem and show how the last three records above are classified, i.e., "Today is 21 Mar.", "Today is 22 Mar.", and "Today is 23 Mar.".

(6 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*	unknown

(2 marks)

c) Formulate the problem for mining association rules like

R1: {2 Trading Day before of HSBC is *Up*}, {Today of CTI is *Level*} → {Next Trading Day of PCCW is Down}

R2: {1 Trading Day before of PCCW is *Up*}, {Today of PCCW is *Down*} → {Next Trading Day of PCCW is Down}.

Show your formulation by writing down the database attributes you will design and give the attribute values of a few sample database records.

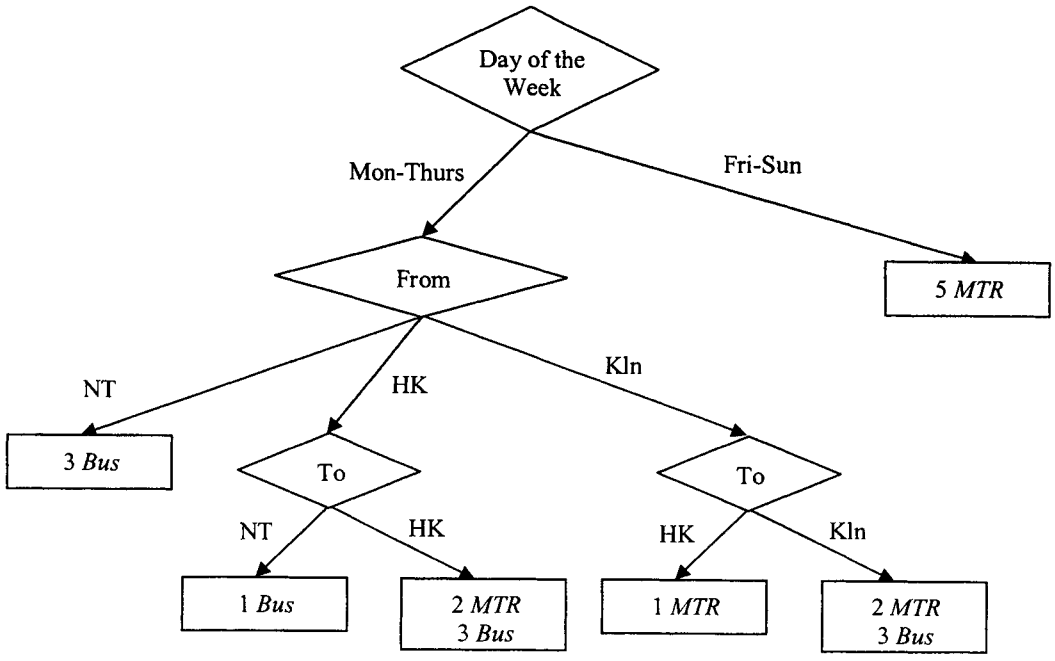
(7 marks)

3. Suppose you are asked by the Octopus Card Company to analyze the following Card Database.

Octopus Card Database

Octopus ID	*Day of the Week*	*From*	*To*	*Transportation Taken*
1	Mon-Thurs	HK	HK	MTR
2	Mon-Thurs	Kln	Kln	Bus
3	Mon-Thurs	NT	Kln	Bus
4	Fri-Sun	HK	HK	MTR
5	Mon-Thurs	Kln	Kln	MTR
6	Mon-Thurs	NT	NT	Bus
7	Mon-Thurs	HK	HK	Bus
8	Fri-Sun	Kln	Kln	MTR
9	Mon-Thurs	Kln	Kln	Bus
10	Mon-Thurs	HK	HK	Bus
11	Fri-Sun	HK	HK	MTR
12	Mon-Thurs	Kln	Kln	Bus
13	Mon-Thurs	HK	HK	Bus
14	Mon-Thurs	Kln	Kln	MTR
15	Fri-Sun	HK	NT	MTR
16	Mon-Thurs	NT	Kln	Bus
17	Mon-Thurs	HK	NT	Bus
18	Mon-Thurs	Kln	HK	MTR
19	Fri-Sun	HK	NT	MTR
20	Mon-Thurs	HK	HK	MTR
21	Mon-Thurs	Kln	Kln	Bus
22	Fri-Sun	Kln	HK	MTR
23	Mon-Thurs	Kln	HK	Bus
24	Mon-Thurs	Kln	HK	Bus
25	Mon-Thurs	HK	HK	MTR

Your project team member tried to construct a decision tree for the first 20 records above and obtained the result as follows.



a) Write down the classification rules that can be extracted from the decision tree above.

(3 marks)

b) Show how the 24th record (Octopus ID=24) should be classified.

(3 marks)

c) Classify the following two records with missing attribute values (denoted *unknown* below).

Octopus ID	*Day of the Week*	*From*	*To*	*Transportation Taken*
100	*Unknown*	HK	HK	?
101	Mon-Thurs	*Unknown*	NT	?

(4 marks)

d) Another project team member commented that the constructed decision tree above is too deep. If the two "*To*" nodes are pruned, what will be the classification rate on the first 20 records of the Octopus Card Database?

(4 marks)

e) If the Octopus Card Company is looking for the most compact set of classification rules, i.e., the smallest decision tree generating the best classification rate, for the training data (the first 20 records), what will the rules be? You may stick with the decision tree provided.

(5 marks)

f) If some more card records are available to update the decision tree above, describe how you carry out the task. You are expected to discuss when reconstruction/update is needed and how reconstruction/update is carried out by giving some typical examples.

(6 marks)

4. Given the following sample Web Page Rating Database containing the HTML source of web pages plus the ratings of different users on these web pages. The web pages are on four separate domains, namely, Finance & Investment, Entertainment & Travel, Education & Lifelong Learning, and Health & Product, and users looked at each web page and indicated on a 3-point scale (Hot; Medium; Cold). There are 50 pages per domain and 1000 users recorded. The HTML web page files can be accessed via their file IDs. As a data miner, you are asked to suggest how data mining techniques can be applied to this database and mine useful knowledge.

Web Page Rating Database

User ID	Domain	HTML File ID	Rating
0001	Finance&Investment	0001	Hot
0001	Finance&Investment	0002	Cold
0001	Finance&Investment	.....	.....
0001	Finance&Investment	0050	Medium
0001	Entertainment&Travel	0051	Hot
0001	Entertainment&Travel	0052	Hot
0001	Entertainment&Travel	.....	.....
0001	Entertainment&Travel	0100	Medium
0001	Education&LifelongLearning	0101	Medium
0001	Education&LifelongLearning	0102	Hot
0001	Education&LifelongLearning	.....	.....
0001	Education&LifelongLearning	0150	Cold
0001	Health&Product	0151	Hot
0001	Health&Product	0152	Hot
0001	Health&Product	.....	
0001	Health&Product	0200	Medium
0002	Entertainment	0001	Hot
.....	.....	.....	.....
0002	Health&Product	0200	Cold
0003	Entertainment	0001	Hot
.....	.....	.....	.....
0003	Health&Product	0200	Hot
.....	.....	.....	.....
.....	.....	.....	.....
1000	Health&Product	0200	Medium

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) Normalize the following medical time series data so that they can be compared or mined fairly.

Medical Time Series Data

Series	Time									
	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
#1	*5.05*	*5.0*	*4.8*	*4.8*	*5.05*	*5.2*	*5.2*	*5.2*	*5.05*	*5.05*
#2	*130*	*130.5*	*130*	*131*	*130.5*	*133*	*134*	*135*	*134*	*134*
#3	*0.8*	*0.82*	*0.82*	*0.81*	*0.81*	*0.8*	*0.8*	*0.78*	*0.78*	*0.78*

(5 marks)

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

[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
9100123	80-120	Male	High	Married	No
9303034	160-200	Female	Medium	Single	Yes
9210126	80-120	Male	Medium	Married	Yes
9142020	120-160	Female	Low	Single	No
9910111	160-200	Male	High	Single	Yes
9576732	80-120	Male	Low	Married	No
9910115		Female	High	Single	Yes
9210120	120-160		Medium		No
9576737	160-200	Female		Married	No

(4 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 such kind of diseases in a hospital/clinic during a period of time and *number_of_case* is the number of such disease cases 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 -