THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF COMPUTING

EXAMINATION

Course : MScIT (61030/88004), MScST (61030/88004), MScIS (61030/61020), MScEC (61030/61027), RS

Subject : COMP5121 Data Mining and Data Warehousing Applications

Group : 101, 102, 103, 104, 1888

Session : 2006 / 2007 Semester I

Date : 14 December 2006 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** questions.
Show your steps and write down any assumption(s) you made.
Standard non-programmable calculator is allowed.

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

COMP5121 Data Mining & Data Warehousing Applications

2006/2007 Fall Examination

Answer ALL questions.

1. In a survey, the following data was collected.
   - □ Among 5000 teenagers who wear jeans,
     - o 3000 play on-line games
     - o 3750 eat chips
     - o 2000 both play on-line games and eat chips
   - □ Among another 5000 teenagers who do not wear jeans,
     - o 3000 play on-line games
     - o 4000 eat chips
     - o 2250 both play on-line games and eat chips

   a) List ALL strong association rules having the form {item1, item2→ eat chips} with support≥15% and confidence≥50%.

   (8 marks)

   b) Compute the interest (lift ratio) of the strong association rules found in part (a).

   (2 marks)

   c) Suppose you are allocated a budget to send out promotion letters of a new chips product to 4000 potential customers. Given a database consisting of 10000 teenager records depicted below, which kind(s) of teenagers, in terms of their attribute values, should be selected if you are asked to make use of the mining results in part (a)? Justify your answer. You may assume that the statistic of the 10000 teenagers' records follows that of the survey result listed above.

Teenager ID	Student Name	Address	Wear Jeans	Play On-line Games	Eat Chips	Other attributes
10001	...	...	Yes	Yes	Unknown	...
10002	...	...	Yes	No	Unknown	...
15000	...	...	Yes	Yes	Unknown	...
15001	...	...	No	No	Unknown	...
15002	...	...	No	Yes	Unknown	...
20000	...	...	No	Yes	Unknown	...

(10 marks)

2. 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 you are asked to provide recommendations (of movie) to customers, describe how you formulate and solve the problem by answering the questions below. Note that you are free to use association rule mining, clustering or classification to accomplish this task.

a) Prepare a task-relevant database based on the records in the two databases above.
(8 marks)

b) Mine the database constructed in part (a) so that the recommendation of "Cinderella Man" to customer 00004 can be determined (i.e., recommend or not recommend). Note that you are NOT required to show all your mining results. Just show your recommendation by referring to the mining results needed.
(12 marks)

3. Consider the following stock price movement data:

Stock Price Database

Stock	Price Movement from 12 December – 24 December										
	12/12	13/12	14/12	15/12	16/12	19/12	20/12	21/12	22/12	23/12	24/12
PCCW	*Up*	*Up*	*Down*	*Down*	*Down*	*Up*	*Up*	*Down*	*Down*	*Up*	*Up*
HSBC	*Down*	*Down*	*Down*	*Up*	*Up*	*Up*	*Down*	*Up*	*Up*	*Down*	*Down*
CTI	*Down*	*Down*	*Up*	*Up*	*Down*	*Down*	*Down*	*Down*	*Down*	*Up*	*Up*
To predict tomorrow's HSBC price movement											
Class	*Down*	*Down*	*Up*	*Up*	*Up*	*Down*	*Up*	*Up*	*Down*	*Down*	

where the movement labels *Up* & *Down* denote the stock price going up & down respectively in the corresponding trading day.

a) Use the ID3 algorithm to construct a decision tree, consisting of one node (i.e. the root node), to predict tomorrow's HSBC price movement. Show your steps.

(10 marks)

b) Based on your solution in part (a), predict 24 December's HSBC price movement.

(4 marks)

c) If one more class attribute is introduced as depicted below, describe how you modify the classifier in part (a) for such a dual classification problem.

Stock Price Database

Stock	Price Movement from 12 December – 24 December										
	12/12	13/12	14/12	15/12	16/12	19/12	20/12	21/12	22/12	23/12	24/12
PCCW	*Up*	*Up*	*Down*	*Down*	*Down*	*Up*	*Up*	*Down*	*Down*	*Up*	*Up*
HSBC	*Down*	*Down*	*Down*	*Up*	*Up*	*Up*	*Down*	*Up*	*Up*	*Down*	*Down*
CTI	*Down*	*Down*	*Up*	*Up*	*Down*	*Down*	*Down*	*Down*	*Down*	*Up*	*Up*
To predict tomorrow's HSBC price movement											
Class	*Down*	*Down*	*Up*	*Up*	*Up*	*Down*	*Up*	*Up*	*Down*	*Down*	
To predict the day after tomorrow's HSBC price movement											
Class	*Down*	*Up*	*Up*	*Up*	*Down*	*Up*	*Up*	*Down*	*Down*		

(6 marks)

4. Given the following transactional database:

Customer	Items Bought
David	30, 50
John	10, 50, 40, 60
Peter	70, 20
Aaron	30, 50, 70

You are asked to cluster them into groups.

a) Propose a distance metric for the given database. Compute and fill in the missing values of the distance matrix below.

$$\begin{array}{r c} & \begin{array}{cccc} \textit{David} & \textit{John} & \textit{Peter} & \textit{Aaron} \end{array} \\ \begin{array}{r} \textit{David} \\ \textit{John} \\ \textit{Peter} \\ \textit{Aaron} \end{array} & \begin{bmatrix} 0 & & & \\ \underline{\hspace{1cm}} & 0 & & \\ \underline{\hspace{1cm}} & \underline{\hspace{1cm}} & 0 & \\ \underline{\hspace{1cm}} & \underline{\hspace{1cm}} & \underline{\hspace{1cm}} & 0 \end{bmatrix} \end{array}$$

(12 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.

(8 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	Risk Level	Marital Status	Disease
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	160-200	Female		Single	Yes
9210120	120-160		Medium		
9576737			Low	Married	No

(6 marks)

b) Suppose you are responsible to design a data warehouse for the hospital authority (HA) and are given three dimensions: (i) doctor, (ii) patient, and (iii) time, and two measures: charge and expense where charge is the fee that the doctor charges a patient for a visit and expense is the cost of the visit calculated by HA.

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

(8 marks)

ii) Assume that the time dimension is characterized by the concept hierarchy L1-day, L2-week, L3-month, L4-quarter and L5-year. The patient dimension is characterized by the concept hierarchy L1-building, L2-district, and L3-region and the doctor dimension is characterized by the concept hierarchy L1-department, L2-hospital, and L3-hospital cluster. What OLAP operations are required to list the total fee collected by the doctors of department D7 in year 2004 if the current data cube is listing the total fee collected by the doctors of hospital H5 in May 2004? Make your own assumption(s).

(6 marks)

- E N D -