THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF COMPUTING

EXAMINATION

Course : MScIT (61030 / 88004) / MScST (61030 / 88004) / PG Scheme / RS

Subject : COMP5121 Data Mining and Data Warehousing Applications

Group : 201, 2011 / 202, 2021 / 205 / 2444, 2888

Session : 2004 / 2005 Semester II

Date : 9 May 2005 Time : 18:30-20:30

Time Allowed : 2 Hours Subject Lecturer : Korris Chung

This question paper has 5 pages (cover included).

Instructions to Candidates :

Answer any **Four** questions. Each question carries equal marks.
Open-book examination.
Show all your steps and write down any assumption(s) you made.

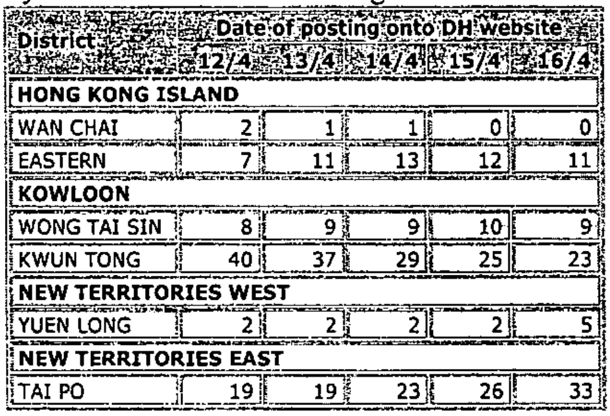
Do not turn this page until you are told to do so !

Answer ANY FOUR questions. Each question carries 25 marks.

1. Sequential pattern mining is typically formulated as finding the frequent subsequences (of itemsets/events) among customers. However in some applications, the problem may be formulated differently so that the frequent subsequences within the same customer can be discovered. As in bioinformatics and stock data mining, it is expected to find the frequent substrings (or subsequences) of a given long string (or sequence). For example, a gene string ACTGACTACAA contains frequent substrings AC with support=3, ACT with support=2, CT with support=2, etc. For the stock price information string UUDLLUUUDD where U, D and L denote the stock price going up, down and level respectively, one may want to find the frequent substring UUD with support=2. Derive an effective mining algorithm that finds the complete set of frequent subsequences according to a given minimum support threshold. Demonstrate how your algorithm works using a small dataset.

(25 marks)

2. Given the following data about the "Number of Buildings with Confirmed SARS Patients by District and Date of Posting".



District	Date of posting onto DH website				
	12/4	13/4	14/4	15/4	16/4
HONG KONG ISLAND					
WAN CHAI	2	1	1	0	0
EASTERN	7	11	13	12	11
KOWLOON					
WONG TAI SIN	8	9	9	10	9
KWUN TONG	40	37	29	25	23
NEW TERRITORIES WEST					
YUEN LONG	2	2	2	2	5
NEW TERRITORIES EAST					
TAI PO	19	19	23	26	33

a) Use the equal depth (frequency) binning method to discretize the numerical values above into 3 intervals.

(5 marks)

b) You are asked to find the associations between districts for the SARS cases. Let minimum support=75% and minimum confidence=100%. Find TWO frequent 3-itemsets and generate the corresponding association rules. Note here that you are NOT required to use the Apriori algorithm to generate the results.

(10 marks)

c) By taking the four taxonomies (i.e. HK Island, Kowloon, NT West and NT East) into considerations, find TWO generalized association rules consisting of at least 1 ancestor items (i.e. HK Island, Kowloon, NT West or NT East) using minimum support=75% and minimum confidence=100%. Again, you are NOT required to use the Apriori algorithm.

(10 marks)

3. a) You are asked to construct a naive Bayesian classifier for the following web page dataset. The first 9 records, i.e. P10-P90, are used for training while the last 3 records, i.e. P100-P120, are used for testing. Compute the classification accuracy for both the training and testing data.

(12 marks)

Web Page ID	A1: No. of outgoing hyperlinks found	A2: No. of incoming hyperlinks found	A3: Depth of Homepage (levels)	Class Attribute: Type of Homepage
P10	Large	Few	Shallow	Hub
P20	Small	Many	Shallow	Ordinary
P30	Small	Many	Deep	Authority
P40	Large	Few	Shallow	Hub
P50	Small	Many	Deep	Ordinary
P60	Medium	Few	Shallow	Authority
P70	Medium	Few	Deep	Hub
P80	Large	Many	Deep	Ordinary
P90	Medium	Many	Deep	Authority
P100	Small	Few	Deep	Ordinary
P110	Medium	Many	Deep	Authority
P120	Large	Many	Shallow	Hub

b) Suppose the decision tree is used to classify a dataset with 3 attributes like the one above where A1 has 3 attribute values, A2 has 2 attribute values, and A3 has 2 attribute values.

i) What is the maximum number of rules that can be generated for such kind of dataset? Briefly describe the characteristics of the dataset for which the maximum number of classification rules will be generated.

(3 marks)

ii) Suppose that the classification rules being formed are restricted to have at most two attribute-value pairs in the antecedent (IF) part of the rules. Briefly describe how it will affect the model construction process and the model usage process. Is there any advantage in doing this? Justify your answer.

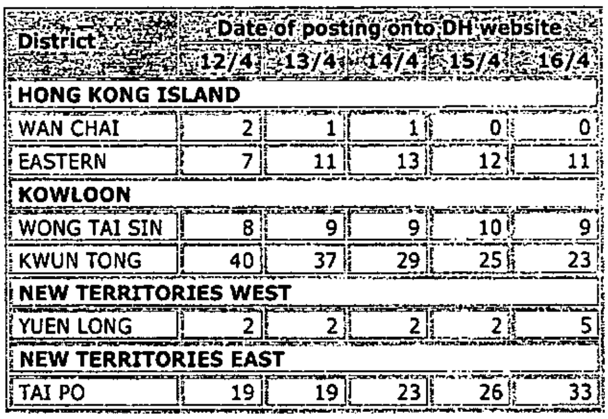
(6 marks)

iii) If only 4 classification rules are obtained for a particular dataset, write down the antecedent (IF) parts of these rules. Note here that the answer is not unique and you are only required to provide one possible set of answers.

(4 marks)

4. a) Given the following data about the "Number of Buildings with Confirmed SARS Patients by District and Date of Posting". You are asked to use the *k*-means algorithm to group the 6 districts into 2 clusters. Assume that WAN CHAI, KWUN TONG and YUEN LONG are initialized as cluster 1 while the remaining districts are assigned to cluster 2. Show the first iteration of the *k*-means clustering process.

(10 marks)



District	Date of posting onto DH website				
	12/4	13/4	14/4	15/4	16/4
HONG KONG ISLAND					
WAN CHAI	2	1	1	0	0
EASTERN	7	11	13	12	11
KOWLOON					
WONG TAI SIN	8	9	9	10	9
KWUN TONG	40	37	29	25	23
NEW TERRITORIES WEST					
YUEN LONG	2	2	2	2	5
NEW TERRITORIES EAST					
TAI PO	19	19	23	26	33

b) Suppose that the *k*-means algorithm is used to group 57 numerical database records into 3 groups. Is it possible to produce two identical clusters that contain the same set of database records, by the *k*-means algorithm? Justify your answer.

(4 marks)

c) If the *k*-means algorithm is used to identify outliers in the dataset, what necessary modifications of the clustering process are required?

(5 marks)

d) The *k*-means algorithm and the agglomerative hierarchical clustering method have their own strength and weakness. Suggest how they can be combined so that certain advantages can be obtained.

(6 marks)

5. a) Suppose you are responsible for designing 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).

(4 marks)

iii) Following part (ii), what OLAP operations are required to list the total fee collected by the doctors of departments D7 & D18 in June and December 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).

(4 marks)

b) Name one corporation/company in Hong Kong which does not already have one (to the best of your knowledge) and should consider starting a data warehouse project for its early and decisive use of information from data. Substantiate your suggestions by describing what data warehouse schema, dimension and fact tables should be formed and used.

(9 marks)

-END-