# THE HONG KONG POLYTECHNIC UNIVERSITY

## DEPARTMENT OF COMPUTING

## EXAMINATION

---

Course : MScIS (61030 / 61020) / MScEC (61030/61027/61028) / RS

Subject : COMP5121 Data Mining & Data Warehousing Applications

Group : 103 / 1031 / 104 / 1041 / 1888

Session : 2004 / 2005 Semester I

Date : 15 December 2004 Time : 18:30 - 20:30

Time Allowed : 2 Hours Subject Lecturer : Korris Chung

---

This question paper has ____6___ pages (attachment included).

---

**Instructions to Candidates :**

Answer **ANY 4** questions.

Open-book examination.
Show your steps and write down any assumption you made.

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

# COMP5121 Data Mining & Data Warehousing Applications

04/05 Fall Term Final Examination

1. Given the following transactional database for sequential association rule mining.

| Customer | Transaction Time | Items |
|---|---|---|
| David | 2 Feb 2002 | 30 |
| David | 10 Feb 2002 | 50 |
| John | 5 Feb 2002 | 10,50 |
| John | 7 Feb 2002 | 50 |
| John | 16 Feb 2002 | 40,60,70 |
| Peter | 16 Feb 2002 | 70 |
| Aaron | 5 Feb 2002 | 10 |
| Aaron | 7 Feb 2002 | 30,50 |
| Aaron | 9 Feb 2002 | 70 |

a) List all the possible sequences arising from customer John.

(5 marks)

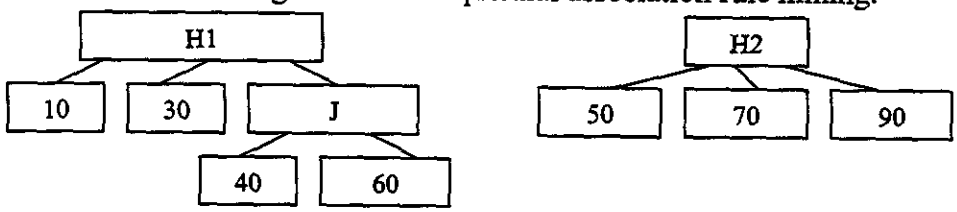
b) Compute the support of the following two sequences:
<(50),(70)>
<(30 50)>
Note: The notation here follows that in the lecture notes.

(4 marks)

c) Let *min_sup*=50% and the modified *AprioriAll* algorithm for handling reoccurrence of itemsets is applied to the transactions above. Find all frequent sequences.

(10 marks)

d) Suppose now the following two taxonomies/hierarchies are associated with the database above for generalized sequential association rule mining.




i) What is the maximum length of the maximal frequent sequences for *min_sup*=X%? Note here that you don't need to apply the AprioriAll algorithm.

(3 marks)

ii) 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*=X%

(3 marks)

2. a) In our term project, we have already applied association analysis to the Microsoft anonymous weblog data. Now, you are assigned another data mining task for this data set, i.e., to cluster the visitor's access patterns, and the following access records are extracted. Recall that there are altogether 294 Vroots being used in Microsoft's website.

| Visitor ID | Vroots visited |
|---|---|
| 20001 | 1008, 1000, 1035, 1016, 1031 |
| 20002 | 1000, 1009, 1016, 1110, 1111 |
| 20003 | 1035, 1115, 1117, 1050, |
| 20004 | 1115, 1117, 1050, 1079, 1088, 1199 |
| 20005 | 1117, 1016, 1050, 1088, 1199 |
| 20006 | 1115, 1088, 1199 |

Here, you may assume that the order information is NOT available, i.e., we do not know whether visitor 20001 has visited Vroot 1008 before Vroot 1000 or not.

i) Propose a distance metric for this clustering task. Compute and fill in the missing values of the distance matrix below.

| | 20001 | 20002 | 20003 | 20004 | 20005 | 20006 |
|---|---|---|---|---|---|---|
| 20001 | 0 | | | | | |
| 20002 | ____ | 0 | | | | |
| 20003 | ____ | ____ | 0 | | | |
| 20004 | ____ | ____ | ____ | 0 | | |
| 20005 | ____ | ____ | ____ | ____ | 0 | |
| 20006 | ____ | ____ | ____ | ____ | ____ | 0 |

(8 marks)

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

(9 marks)

b) Weblog mining should be able to discover potentially useful pattern/knowledge. According to the web master, your company's web server registers every access of the web page as a record in the Weblog database which includes the URL requested, the IP address from which the request originated, and a timestamp with format *yy-mm-dd|hh:mm:ss* corresponding to year, month, day, hour, minute and second respectively.

i) Briefly describe the typical steps in a KDD process (e.g., cleaning) and how they are implemented in our weblog mining context.

(4 marks)

ii) How will the weblog mining and KDD process improve the company's services? List four examples.

(4 marks)

3. You are working for the MONDAY telecom company and are given some customer records. Your manager asks you to find the classification rule(s) for high and low usage customers. The data are given below.

| 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 |

a) Suppose you take use of the naive Bayesian classification method to solve the problem. Show how the following customer record should be classified.

| Customer ID | Monthly Income | Age | Education | Marital Status |
|---|---|---|---|---|
| 9100100 | Medium | Young | University | Married |

(10 marks)

b) Following part (a). Show how the following customer record with missing value should be classified.

| Customer ID | Monthly Income | Age | Education | Marital Status |
|---|---|---|---|---|
| 9100101 | Medium | | University | Married |

(5 marks)

c) Suppose now Usage is categorized as High, Medium, and Low and some records with Medium Usage are added. Do you think that the naive Bayesian classification method can still be used? Justify your answer.

(4 marks)

d) What are the advantages of naive Bayesian classification compared with decision tree induction?

(6 marks)

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

| 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 | Many | Few | 7 | Hub |
| P20 | | Few | 2 | Ordinary |
| P30 | Few | Many | | Ordinary |
| P40 | Many | | 10 | Hub |
| P50 | Few | Many | 3 | Ordinary |
| P60 | Medium | Few | | Hub |
| P70 | Many | | 7 | Ordinary |
| P80 | | Many | 10 | Hub |

(6 marks)

b) Why data cleaning and data preprocessing are important in knowledge discovery in databases? Why they are so time consuming?

(6 marks)

c) In our FAQ on data warehousing, we can see the possibility of the MTR Corporation to start a data warehouse for its Octopus data. Name one corporation/company, e.g. PCCW or Jockey Club, which should consider starting a data warehouse for its early and decisive use of information from data. Substantiate your suggestion by describing what data warehouse schema, dimension and fact tables should be formed and used.

(13 marks)

5. As a Hong Kong citizen who loves Hong Kong, you want to apply what you have learnt from COMP5121 to discover knowledge that might be useful to combat the SARS (Severe Acute Respiratory Syndrome) problem when it comes back. You have extracted the data about the "Number of Buildings with Confirmed SARS Patients by District and Date of Posting" from the Department of Health's website as shown in Attachment A.

   a) If the classification technology is adopted, describe how you formulate and solve the problem. You are recommended to limit your discussions to the following related issues and make use of the data in Attachment A to exemplify your solutions. You may assume that the "number of buildings" in numerical values can be converted into categorical labels like "many buildings", "some buildings" and "less buildings".
      i) What is the class attribute? What are the ordinary attributes for classification?
      ii) What classification model(s) will you suggest?
      iii) What (hypothetical) classification knowledge do you expect?

      (13 marks)

   b) If the clustering technology is adopted, describe how you formulate and solve the problem. Again, you are recommended to limit your discussions to the following related issues and make use of the data in Attachment A to exemplify your solutions.
      i) What database attributes will you use for clustering?
      ii) What clustering technique(s) will you suggest?
      iii) What (hypothetical) clustering knowledge do you expect?

      (12 marks)

- END -

## Attachment A. 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 | 17/4 | 18/4 | 19/4 | 20/4 | 21/4 | 22/4 | 23/4 | 24/4 | 25/4 | 26/4 | 27/4 | 28/4 | 29/4 | 30/4 |
| **HONG KONG ISLAND** | | | | | | | | | | | | | | | | | | | |
| CENTRAL AND WESTERN | 1 | 0 | 0 | 0 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| WAN CHAI | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 0 | 0 | 1 | 1 |
| EASTERN | 7 | 11 | 13 | 12 | 11 | 9 | 9 | 8 | 7 | 7 | 8 | 8 | 5 | 3 | 4 | 4 | 5 | 5 | 5 |
| SOUTHERN | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **KOWLOON** | | | | | | | | | | | | | | | | | | | |
| YAU TSIM MONG | 7 | 10 | 9 | 6 | 6 | 5 | 4 | 2 | 2 | 1 | 1 | 4 | 5 | 4 | 3 | 2 | 2 | 3 | 3 |
| SHAM SHUI PO | 8 | 12 | 12 | 9 | 8 | 9 | 11 | 12 | 8 | 7 | 5 | 5 | 7 | 7 | 5 | 6 | 7 | 7 | 6 |
| KOWLOON CITY | 9 | 10 | 11 | 13 | 13 | 11 | 7 | 4 | 5 | 4 | 2 | 4 | 4 | 4 | 5 | 2 | 1 | 1 | 2 |
| WONG TAI SIN | 8 | 9 | 9 | 10 | 9 | 8 | 6 | 6 | 5 | 7 | 6 | 8 | 10 | 11 | 10 | 9 | 8 | 8 | 6 |
| KWUN TONG | 40 | 37 | 29 | 25 | 23 | 24 | 21 | 17 | 16 | 13 | 13 | 15 | 14 | 13 | 12 | 11 | 11 | 7 | 7 |
| **NEW TERRITORIES WEST** | | | | | | | | | | | | | | | | | | | |
| KWAI TSING | 15 | 15 | 16 | 17 | 17 | 15 | 11 | 15 | 14 | 13 | 9 | 12 | 14 | 13 | 11 | 10 | 10 | 12 | 12 |
| TSUEN WAN | 4 | 3 | 5 | 4 | 4 | 6 | 6 | 4 | 3 | 2 | 3 | 4 | 5 | 4 | 4 | 3 | 3 | 3 | 3 |
| TUEN MUN | 12 | 10 | 9 | 7 | 10 | 6 | 4 | 3 | 3 | 1 | 1 | 3 | 3 | 2 | 1 | 3 | 4 | 4 | 4 |
| YUEN LONG | 2 | 2 | 2 | 2 | 5 | 6 | 6 | 7 | 6 | 6 | 6 | 7 | 6 | 6 | 6 | 7 | 8 | 9 | 9 |
| **NEW TERRITORIES EAST** | | | | | | | | | | | | | | | | | | | |
| NORTH | 2 | 3 | 4 | 4 | 4 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 4 | 3 | 2 | 3 | 4 | 3 | 4 |
| TAI PO | 19 | 19 | 23 | 26 | 33 | 31 | 27 | 22 | 24 | 21 | 24 | 21 | 23 | 28 | 29 | 25 | 21 | 15 | 16 |
| SHA TIN | 22 | 23 | 22 | 16 | 19 | 19 | 20 | 19 | 18 | 18 | 18 | 20 | 19 | 15 | 12 | 11 | 9 | 8 | 4 |
| SAI KUNG | 8 | 8 | 10 | 7 | 7 | 7 | 8 | 7 | 6 | 7 | 6 | 5 | 5 | 4 | 4 | 3 | 2 | 1 | 1 |
| ISLANDS | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |