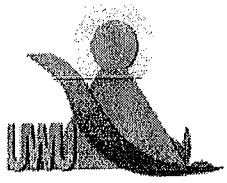

UVA WELLASSA UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE & TECHNOLOGY

END SEMESTER EXAMINATION – SEMESTER I – 2009/2010

CST412-3 Data Mining and Data Warehousing

Time allowed: 3 hours

This paper has 6 questions. Answer 5 questions. And attach your question paper to the answer script

Q1.

a. What is meant by data mining? [3 marks]

b. Define any three of the following data mining functionalities and describe each of them using suitable examples.

i. Characterization
ii. Discrimination
iii. Classification
iv. Prediction
v. Clustering
vi. Evaluation Analysis [9 marks]

c. Briefly explain the steps in the process of knowledge discovery in a data mining system using a suitable example. [8 marks]

Q2.

a. What are the main differences between operational database systems and data warehouses? [4 marks]

b. 'Surveys indicate that, the use of Data Mining and Data Warehousing technology is becoming increasingly popular in the industry at present.' Explain the reason for this trend. Your answer should include a comparison with traditional database management systems. [8 marks]

c. Discuss, at least two possible areas in Sri Lanka where data mining and data warehousing technology can be applied for the decision making process. A justification of the arguments presented in the answer is required. [8 marks]

Q3.

a. Discuss the importance of Pre-processing of data in data mining and data warehousing industry.

b. Consider the following student marks that are in ascending order:

0,23,32,39,39,40,40,43,45,45,49,52,52,52,60,61,61,69,71,71,71,71,85,92,92,99

i. To smooth the data, the following two data pre-processing methods are to be applied on the above data set.

- Bin means
- Bin boundaries

Assuming number of bins are 4, discuss the bin mean and bin boundaries.

ii. Using the above data set, briefly describe how one could determine the outliers.

c. Apply Min-max normalization to transform the value 52 for marks onto the range [0.0, 1.0] using the data set given in Part (b) above.

d. Apply Z-score normalization to transform the value 52 for marks, where the standard deviation of marks is 22.64 using the data set given in Part (b) above.

Note:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A.$$

$$v' = \frac{v - mean_A}{stand\_dev_A} \qquad v' = \frac{v}{10^j},$$

[20 marks]

Q4.

a. What are the UML Diagrams available in object oriented design? Define three of them. [6 marks]

b. Adwel is a small advertising company. The company wishes to develop a software system to provide facilities such as on-line access to customers to place their advertising requests and the company to manage the advertising process. The administrative staff of the company is responsible for the management of advertising campaigns. An advertising campaign required by a customer may compose of several advertisements. An advertisement may be a TV advertisement, news paper advertisement or a web-site advertisement.

The administrative staffs create a campaign and allocate technical staff according to their expertise using the system. The technical staff can use the system to update campaigns according to the progress. The administrative staff monitors the progress of campaigns. Once the campaign is completed the administrative staff performs the billing operation to charge the customer.

i. Identify the classes in this system and draw a class diagram. [14 marks]

Q5. Suppose that a warehouse named 'University Admissions' consists of the tree dimensions location, time and course and the three measures total marks, number of students and GPA (Grade Point Average) where GPA is based on individual paper.

a. Draw a schema diagram for the above Data Warehouse.

b. Discuss the OLAP operations in the Multidimensional Model such as Roll up, Drill Down, Slice, Dice and Pivot.

c. Define the DMQL for the following

i. Define cube for the University Admission

ii. Define dimension for location, time, and courses [20 marks]

Q6.

a. Compare features between OLAP (On Line Analytical Processing) and OLTP (On Line Transactional Processing) Systems.

b. What is a cube in a data warehouse? Explain how it is constructed and used.

c. The following tables show the selected information of academic year, district name, and course of study and cut-off Z-score for the university admission.

	Year = 2005			Year = 2006		
	Colombo	Gampaha	Kalutara	Colombo	Gampaha	Kalutara
Medicine	2.03	1.8222	1.8612	2.4521	1.8121	1.8721
Dental Surgery	1.9876	1.8128	1.8554	1.8521	1.8021	1.8621
Veterinary Science	1.8983	1.7624	1.8125	1.8321	1.7521	1.8021
Agriculture	1.2955	1.2022	1.2064	1.8212	1.3251	1.3201

Engineering (MPR)	1.9937	1.7183	1.8284	2.0125	1.8521	1.632
Engineering (EM)	1.8385	1.6468	1.798	1.8984	1.521	1.6214
Engineering (TM)	1.9305	1.6937	1.8521	1.7021	1.4951	1.5321
Quantity Surveying	1.7674	1.5905	1.7271	1.7231	1.3821	1.4201
Computer Science and Technology	1.523	1.3456	1.4235	1.3456	1.4321	1.3456
Management	1.6521	1.5856	1.5781	1.6821	1.6021	1.5321
Commerce	1.6068	1.5652	1.5621	1.6201	1.5121	1.5112
Arts	1.253	1.2919	1.3334	1.2521	1.3201	1.4352
Arts (SP) – Mass Media	1.2081	1.2081	1.2081	1.2321	1.2531	1.3201

Answer the following using above data set

i. Perform Roll-up, drill down, slice and dice operations.

ii. Draw a suitable star schema diagram to represent the above enterprise-wide data mart.

iii. Write down the cube definition and the dimension definition for each of the attributes in the proposed schema diagram in part (b) (11) above. State any assumptions made.

[20 marks]