M.Sc. TE (2 Yr) 2nd Semester 20 February 2023 (Afternoon)

47

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

MID SEMESTER EXAMINATION **SUMMER SEMESTER, 2021-2022**
DURATION: 1 HOUR 30 MINUTES **FULL MARKS: 75**

CSE 6293: Data Warehousing and Mining

Programmable calculators are not allowed. Do not write anything on the question paper.

Answer **all 3 (three)** questions. Figures in the right margin indicate full marks of questions.

1. a) Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. 3+5+5 +5

   i. Mention three classes of schemas that are popularly used for modeling data warehouses.

   ii. Draw a schema diagram for the above data warehouse using one of the schema classes that you listed for question 1.a)i.

   iii. Starting with the base cuboid [*day, doctor, patient*], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2020?

   iv. To obtain the same list, write an SQL query assuming the data is stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge).

   b) *Bitmap indexing* is useful in data warehousing. Mention the pros and cons of using a bitmap index structure. 7

2. a) Discuss the differences and similarities of data warehouse and database. 7

   b) Mention the steps involved in data mining when viewed as a process of knowledge discovery. Why *concept hierarchies* are useful in data mining? 8

   c) Formulate an example where data mining is crucial to the success of a business. What data mining functionalities does this business need (e.g., assume the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis? 10

3. a) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. 5+5+ 2

   i. Use *smoothing by bin means* to smooth the above data, where the bin depth is 3. Illustrate your steps.

   ii. How would you determine outliers in the data?

   iii. What other methods are there for data smoothing?

   b) What are the benefits of applying dimensionality reduction to a dataset? Why does PCA perform linear orthogonal transformation on the data? 8

   c) Suppose that **x = (5, 0, 3, 0, 2, 0, 0, 2, 0,0)** and **y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)** are the two term-frequency vectors. Compute the cosine similarity between the two vectors. 5