Islamic University of Technology

Organisation of Islamic Cooperation (OIC)

Department of Computer Science and Engineering (CSE)

MID SEMESTER EXAMINATION **WINTER SEMESTER, 2018-2019**

Duration: 1 Hour 30 Minutes **Full Marks: 75**

CSE 6249: Data Warehousing and Mining

Programmable calculators are not allowed. Do not write anything on the question paper. There are **4(four)** questions. Answer any **3 (three)** of them. Figures in the right margin indicate marks.

---

1. (a) What is the basic difference between database driven application and data-mining application? Briefly present one application scenario of a data-mining application that can be used for the people and society around you. [5]

   (b) There are a number of methods to deal with missing values in data preprocessing phase in data-mining applications. Describe them. [5]

   (c) With appropriate argument, derive the formula for proximity measures for binary attributes. You need to explain particularly why the formulas are slightly different for symmetric binary attributes and asymmetric binary attributes. Place suitable example data to establish your argument. [10]

   (d) What are the major challenges of mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of data (such as a few hundred tuple data set)? [5]

2. (a) Define the following terms with examples: [4]
   i. Data objects and attributes
   ii. Ordinal attribute
   iii. Interval-based attribute and
   iv. Ratio-scaled attribute

   (b) Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): [6]
   i. Compute the Euclidean distance between the two objects.
   ii. Compute the Manhattan distance between the two objects.
   iii. Compute the Minkowski distance between the two objects, using q = 3.
   iv. Compute the Supremum distance between the two objects.

   (c) We can perform on-line analytical processing directly in an operational database instead of building a separate data warehouse. Present arguments to strengthen the statement. Also provide logics to weaken it. Finally justify your position. [10]

   (d) In data mining the data set may contain hundreds of attributes many of which are irrelevant or redundant. Propose a general outline to minimize the problem. [5]

3. (a) Euclidean, Manhattan and Minkowski distances have a major limitation. Briefly explain. Also explain how cosine similarity eliminates it. [5]

   (b) Consider the following Term-Frequency Vector between document 1 and document 2: [10]

Table 1: Term Frequency Vector for question no. 3(b)

Doc No.	Term Frequency						
..	Coach	Win	Goal	Draw	Season	Best Player	Penalty
Document 1	7	3	1	2	0	0	1
Document 2	0	1	4	0	2	1	0

Based on the document property select a suitable measure of similarity. Finally calculate to what extent they are similar.

(c) Present the definition of data warehouse given by William H. Inmon. The definition leads to a number of key components of data warehouse. Briefly discuss them. [5]

(d) Explain the concept of Entity Identification Problem with a suitable example. [5]

4. (a) Why normalization is used in data mining algorithms? Use the following data to normalize them as directed: [5]

200, 300, 400, 600, 1000

i. Min-max normalization by setting min = 0 and max = 1

ii. z-score normalization

iii. Normalization by decimal scaling

(b) Explain the concept of Snowflake and Star schema with suitable example. Also highlight the strength and weakness of each model. [10]

(c) Suppose that the Statistical Bureau of Bangladesh (SBB) wants to build its own data warehouse for a number of purposes. [10]

SBB is interested to analyze the correlation between the followings:

- Income of people and their geographic location
- Children education level and their financial status
- Results of Higher Secondary Examination and colleges

i. Determine the major dimensions and measures for the given scenario.

ii. Draw a start schema diagram for the data warehouse.

iii. Add other features so that you can convert the start schema into an equivalent snowflake schema.