M.Sc. Engg./Ph.D. CSE 194 16 March 2018 (Afternoon)

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)
ORGANISATION OF ISLAMIC COOPERATION (OIC)
Department of Computer Science and Engineering (CSE)

MID SEMESTER EXAMINATION **WINTER SEMESTER, 2017-2018**
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 transactional data? Explain with example. 4
   b) 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. 6
   c) Briefly explain the different types of OLAP operations with suitable diagram. For each operation also present an example of its equivalent SQL statement. 10
   d) *"Boxplots are a popular way of visualizing data distribution"*. First define boxplot and then place an example to justify the argument. 5

2. a) Define Data Objects and Attributes. What is nominal attribute? *"Mean, median and mode of nominal attribute data have no meaningful interpretation."* Justify your position with suitable example. 8
   b) Explain the main difference between ordinal and interval-scaled attribute. 3
   c) Although the mean is the singlemost useful quantity for describing a data set, it is not always the best way of measuring the center of the data. A major problem with the mean is its sensitivity to extreme (e.g., outlier) values. Even a small number of extreme values can corrupt the mean. Place example to explain this fact. To reduce this problem "trimmed mean" is used. Explain it. 6
   d) Given two objects represented by the tuples (12, 4, 42, 10) and (20, 2, 36, 8): 8

      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. iv. Compute the supremum distance between the two objects.

3. a) What is the purpose of using *Jaccard coefficient*? Place example in this regard. 5
   b) Both Manhattan and Euclidean distance satisfy a number of mathematical properties. Briefly mention them. 7
   c) *"Traditional distance measures do not work well for sparse numeric data."*- Justify with a suitable example. Also propose and explain a suitable measure to handle such data. 8
   d) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. 5

4. a) What is the main purpose of normalization? Define the following normalization methods with examles? 8

      i. min-max normalization.
      ii. z-score normalization.
      iii. z-score normalization using the mean absolute deviation instead of standard deviation.
      iv. normalization by decimal scaling.

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

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 or equivalent and College

i. Explain and draw a star schema diagram for the data warehouse.
ii. Also propose the snowflake schema diagram for the same data warehouse.
iii. Finally highlight the comparative strength and weakness of both approaches.

c) Measures can be organized into three categories such as distributive, algebraic and holistic. Briefly explain them. 7