CITS3401 Sample Midterm Questions with Sample Answers

Section I: Data Warehousing

Short Answers

1. (5 marks) What is a data cube? Use an example to explain the OLAP operations roll up and drill down in relation to a data cube.

**Answer:** A data cube is a **lattice of cuboids**. It is often organised around a central theme, such as “road crash fatalities”. This allows data to be modelled and viewed in multiple dimensions. For example, in the crashes data set, if we pick “day of crashes”, “state” and “road user”, we can roll-up from the **base cuboid** (day of crashes, state, road user) to **2D cuboids** (day of crashes, state) by aggregating over the different types of users. This roll up process can continue to the **apex cuboid** to find the overall fatalities. The drill down process is the opposite direction, starting from the apex cuboid to the base cuboid.

(NOTE: Better to draw a lattice to help illustration) The figure below illustrates the various level of cuboids, where A can represent “day of crashes”, B ”state” and C “road user”:

![Diagram of cuboids](image)

2. (5 marks) Explain the concept of a data warehouse and the main steps required for constructing a data warehouse.

**Answer:** A data warehouse is often organised around major subjects, such as “crashes”, “road users” and etc. **(Subject-Oriented)**; it focuses on modelling and analysing data for decision makers, not for daily operations or transaction processing (OLAP vs. OLTP); often constructed by integrating multiple heterogeneous data sources **(Integrated)**; the time horizon is often longer (e.g. 5-10 years of historical data) than that of operational systems **(Time Variant)** and presents as a physically separate store of data transformed from the operational environment **(Non-Volatile)**.

(NOTE: Answers contain the right keywords, but without brief explanation will receive partial marks.)

The major steps are:

Choose a **business process** to model; decide on the **atomic level of data** for this chosen business process; choose the **dimensions** that will apply to each fact table record, choose the **measures**.

(NOTE: A list of steps summarising how you completed your project is also acceptable.)

3. (5 marks) Explain the meaning of star schema and snowflake schema in relation to a data warehouse.

**Answer:** A star schema consists of a fact table in the middle connected to a set of dimension tables.

For snowflake schema, some dimensional hierarchy is normalised into a set of smaller dimension tables, forming a shape similar to snowflake.

(NOTE: example illustrations are also acceptable.)

4. (5 marks) Explain what multiway array aggregation means and how it achieves simultaneous aggregation.

**Answer:** Multiway Array Aggregation separate the cube into **equal sized chunks** of the same dimensions. These multi-dimensional chunks are **given direct addresses**, such as 1 to 64. Then the calculation of the low dimensional chunks are carried out according to the address order. The address order will save memory if the smallest plane is chosen for the ordering. **The computation of each chunk will contribute to a portion of the calculations of each plane (i.e. lower dimension cubes).** That way, the calculation is shared among all parent cuboids, achieving simultaneous aggregation, avoiding visiting any data points twice.
5. (5 marks) Explain the difference between distributive, algebraic and holistic measures, using example measures.

Answer: Distributive measures are functions that can be computed in a distributed manner by applying the same function on partitioned sets. For example, count(), min() and max() are distributive.

Algebraic measures are functions that can be computed by an algebraic function of arguments that can be computed in a distributive manner. For example: average() and standard_deviation().

Holistic function are functions that cannot be computed in a distributive or algebraic manner. For example, median(), mode() and rank.

Long Answers

(20 marks)
Suppose that a data warehouse consists of three dimensions time, doctor and patient, and two measures count (the number of patients examined) and charge (fee that a doctor charges a patient for a visit).

1. Draw either a star or a snowflake schema for the above data warehouse.

2. 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 2010?

   Answer: This query can be realised through a roll-up operation to create a 2D cuboid of [day, doctor], and aggregate on the patient dimension (or you can say group by day and the doctor) for measure “fee”.

   (NOTE: Optional - you can draw a STARNET to indicate the query footprint).

3. Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee paid by patient John Citizen in the years 2009 and 2010 combined?

   Answer: This query can be realised through a roll-up operation to create a 2D cuboid of [patient, day], and aggregate on the year dimension (or you can say group by day and patient) for measure “fee”. To find 2009 and 2010 combined, in the concept hierarchy for year, you should create a consolidated element that contains both 2009 and 2010.

   (NOTE: Optional - you can draw a STARNET to indicate the query footprint).

Section II: Data Exploration, Data Cleaning and Data Reduction

Short Answers

1. (5 marks) What are the common strategies for dealing with missing data?
Answer:
- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant: e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as regression, Bayesian formula or decision tree

(NOTE: for filling in automatically, manage two of the four will get full marks)

2. (5 marks) What is a 5 number summary of the dataset? How is it related to a boxplot?

   ![Box Plot Graph](image)

   (NOTE: No need to point out the outliers.)

3. (5 marks) How would you classify the various data reduction techniques we discussed in the lectures?

   - Dimensionality Reduction:
     - Mapping or Projecting on to an efficient feature space.
     - Wavelet Transforms, Principal Component Analysis
     - Attribute Subset Selection
   - Numerosity reduction:
     - Parametric methods: a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data. (Outliers may also be stored.)
     - Nonparametric methods for storing reduced representations of the data include
       - Regression Models, Histograms, clustering
   - Data Compression
     - Lossless vs. lossy compression

   (NOTE: Alternatively, one can classify by whether the methods are aiming at reducing attributes, instances or both. Any 5 important techniques briefly discussed would be acceptable.)

Long Answers

(10 marks) Given the following data table and the formulas, demonstrate the process of calculating correlation for Table 1 and chi-square for Table 2.
<table>
<thead>
<tr>
<th>Time Point</th>
<th>AllElectronics</th>
<th>HighTech</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>t2</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>t3</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>t4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>t5</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 1

<table>
<thead>
<tr>
<th>Student</th>
<th>Read Science Fiction</th>
<th>Play Chess</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>s2</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>s3</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>s4</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>s5</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 2

Table 1:

\[ \bar{A} = \frac{6 + 5 + 4 + 3 + 2}{5} = $4 \]

\[ \bar{B} = \frac{20 + 10 + 14 + 5 + 5}{5} = $10.80 \]

\[ \sigma_A = \frac{1}{\sqrt{5}} \left[ (6 - 4)^2 + (5 - 4)^2 + (4 - 4)^2 + (3 - 4)^2 + (2 - 4)^2 \right] \]

Similarly, you can work out \( \sigma_B \). N is equal to 5. Then

\[ r_{A,B} = \frac{[6*20 + 5*10 + 4*14 + 3*5 + 2*5] - 5*4*10.80}{5 * \sigma_A * \sigma_B} \]

(NOTE: You do not need to work out the exact value, correct procedure will get full marks).

Table 2:

First, you need to create a contingency table:

<table>
<thead>
<tr>
<th></th>
<th>Science Fiction</th>
<th>No Science Fiction</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chess</td>
<td>2 (3*3/5)</td>
<td>1 (3*2/5)</td>
<td>3</td>
</tr>
<tr>
<td>No Chess</td>
<td>1 (2*3/5)</td>
<td>1 (2*2/5)</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

The values in the brackets are \( e_{ij} \)

(NOTE: the above contingency table with \( e_{ij} \) calculated will receive full marks, no need to work out the final value).