CSCE 566, Assignment 2

Dr. Vijay Raghavan

Assigned: March 2, 2016

Due: March 14, 2016 (First 3 questions), April 4 (Question 4)

Note:

1. You must show all details of work for each question

2. Staple the question and answer sheet together

3. Make a cover with Name, CLID

4. Number all pages and give an index to each question

5. Any kind of cheating will NOT be tolerated. More information can be found on class Web page on cheating policy.

Question 1:

A Database has six transactions. Let min_sup=50% and min_conf=60% [20 points]

T_id	Items
1	A, B, C
2	A, C, D, E
3	A, B, C, D, E
4	B, D, E
5	A, B, C, E
6	A, B, C, D

a. Find all frequent Itemsets using Apriori algorithm.

b. List all of the strong association rules (with support s and confidence c) matching the following meta-rule, where X is a variable representing customers, and item-I denotes variables representing items (e.g., "A", "B", etc):

   For all X with a transaction in the database, Buys(X, item-n1) ∧ buys(X, item-n2) => buys(X, item-n3) [s, c]

   Note: To find only rules fitting this meta-rule, itemsets with size > 3 are not needed.

Question 2:

Suppose that a data warehouse for an Airline consists of the four dimensions origin, destination, flight type, and day, and two measures, revenue and number of passengers. The flight type has three levels "flight_id < flight_series < flight_family". Flight_family has three values: Airbus, Boeing or McDonnell Douglas. For origin and destination, the concept hierarchy involves "airport < city < state < country < region" and for day, the hierarchy consists of "day < month < quarter < year." [20 points]

a) Draw a star schema for the data warehouse.

b) Starting from the base cuboid, what specific OLAP operations should one perform in order to list the total number of passengers that flew in Boeing 767 series from California to Europe in 2011?

c) Suppose there are four cuboids materialized:

cuboid 1: {month, flight_family, origin_city}

cuboid 2: {year, flight_id, origin_state}

cuboid 3: {year, flight series, origin_state, destination_country}

cuboid 4: {flight_id, origin_state, destination_country} where year =2011

Which of the above cuboids would you select for the query in part b)? Explain your reasons.

d) How many cuboids does this data warehouse contain, including base and apex cuboids?

Question 3:

Suppose that the given table is derived by attribute-oriented induction. [20 points]

1. Transform the table into a crosstab showing the associated t-weights and d-weights.
2. Map the class McDonalds into a (bidirectional) quantitative descriptive rule with attribute location, for example, ∀ X, McDonalds (X) ⇔ (Location(X) = "USA" ∧ Years(X) = "2000 - 2010") [t : x%, d :y%]…∨⊖ (…)[t : w%, d : z%].
3. Map the class Subway into a (bidirectional) quantitative descriptive rule with attribute location, for example, ∀ X, Subway (X) ⇔ (Location(X) = "USA" ∧ Years(X) = "2000 - 2010")[t : x%, d :y%]…∨⊖ (…)[t : w%, d : z%].

Class	Location	Years	Number of new locations
McDonalds	USA	Start -2000	8123
		2001-2010	4681
	Canada	Start -2000	936
		2001-2010	218
	Rest of the World	Start -2000	5454
		2001-2010	6251
Subway	USA	Start -2000	11782
		2001-2010	7685
	Canada	Start -2000	1391
		2001-2010	702
	Rest of the World	Start -2000	981
		2001-2010	2372

Question 4:

Implement a star schema for the sales example in the text book (Page Number 140 in 3rd edition, Page 115 in 2nd edition) using the given data. [40 points]

The data is available on the class webpage under the Assignment section.

http://www.cacs.louisiana.edu/~cmps566/assignment2_data.html

Design a simple user interface that allows you to perform roll up and drill down operations by each dimension. Your interface should allow the specification of a particular aggregate level of the cuboid, and display the number of units sold and dollars earned for every tuple at that aggregate level.

The user should able to select any on dimension and indicate, for the selected dimension, the concept hierarchy level to which either a drill-down or a roll-up operation should be performed with respect to current cuboid and obtain the measures associated with each tuple of the newly specified cuboid.

In case the output desired has more than 2 dimensions, the user should be able to specify which dimensions are in row order, column order and page order, respectively. In case of page or row order has more than 2 dimensions, specify the order (inner-most to outer-most) and the specific value or member of interest with respect to each page order dimension.

Hints

- *Use SQL statements to implement each roll-up, drill-down*
- *All the calculation of the response of each OLAP operation (SQL statements) should be done from the base cuboid.*
- *See Example 3.7 in 2nd edition of the text book for reference.*

You can implement your design using any programming language (i.e., C++, PERL, Java, VB, C#)

Deliverables

- Source code.
- Demo.