CS412: An Introduction to Data Warehousing and Data Mining **Fall 2016**

Assignment 2

Po-Wei Chan *Released Date: October 5, 2016*

Question 1

(24') Assume a base cuboid of 10 dimensions contains only two base cells:

$$(1)\ (a_1, a_2, a_3, a_4, b_5, ..., b_9, b_{10}), \text{ and } (2)\ (b_1, b_2, b_3, b_4, b_5, ..., b_9, b_{10}),$$

where a_i is not equal to b_i (for any $i = 1, ..., 10$). The measure of the cube is count. Answer the following questions.

a. **How many nonempty aggregated (i.e., non-base) cells does a complete cube contain?**
Answer: 1982. **Hint:** For the cuboids with at least one of the first four dimensions being not aggregated (i.e., not $*$), like cuboid $(D_1, D_2, *, *, ..., *)$ or $(D_1, D_2, *, *, D_5, D_6, ..., D_{10})$, there are two nonempty cells, corresponding to two different base cells. There are $(2^4 - 1) \cdot 2^6 = 960$ such cuboids. For the cuboids with all of the four dimensions being aggregated, like cuboid $(*, *, *, *, D_5, D_6, ..., D_{10})$, there is only one nonempty cell. There are $2^6 = 64$ such cuboids. Therefore, the total number of nonempty **aggregated** cells is $960 * 2 + 64 - 2 = 1982$.

b. **How many nonempty aggregated cells does an iceberg cube contain, if the condition of the iceberg cube is** $count \geq 2$**?**
Answer: 64. **Hint:** To have $count \geq 2$, the first four dimensions have to be aggregated (i.e., $*$) while the last six dimensions being either b_i or $*$. There are $2^6 = 64$ such cells.

c. **How many non-star dimensions does the closed cell with count 2 have?**
Answer: 6. **Hint:** The closed cell with $count = 2$ is $(*, *, *, *, b_5, b_6, ..., b_{10})$.

d. **How many closed cells are there in the full cube?**
Answer: 3. **Hint:** The three closed cells are $(*, *, *, *, b_5, ..., b_{10})$, $(a_1, a_2, a_3, a_4, b_5, ..., b_{10})$, and $(b_1, b_2, b_3, b_4, b_5, ..., b_{10})$.

Question 2

(36') Given the dataset, it contains 50 rows, with each row representing a business. For each business, there are six fields (BusinessID, City, State, Category, Rating, Price). The fields are separated by tabs. We now want to construct a cube over four given dimensions (Location, Category, Rating, Price) with "Count" as the measure. Note that in the Location dimension, there is a concept hierarchy, i.e., City and State. Based on the dataset, answer the following questions.

a. **How many cuboids are there in this cube?**
Answer: 24. **Hint:** Suppose the i^{th} dimension has L_i level, the number of cuboids would be $\prod_i (L_i + 1)$. The number here is $(2 + 1) * (1 + 1) * (1 + 1) * (1 + 1) = 24$.

b. **How many nonempty cells are there in the cuboid (Location(City), Category, Rating, Price)?**
Answer: 46. **Hint:** As long as you understand the concept of cuboid and cell, you can find the answer of problems (b.) to (f.) easily by writing a simple code. They are basically counting numbers only.

c. **Drill up by climbing up in the Location dimension from City to State. How many nonempty cells are there in the cuboid (Location(State), Category, Rating, Price)?**
Answer: 36.

d. **How many nonempty cells are there in the cuboid (*, Category, Rating, Price)?**
Answer: 22.

e. **What is the count for the cell (Location(State) = "Illinois", *, Rating = "3", Price = "moderate")?**
Answer: 5.

f. **What is the count for the cell (Location(City) = "Chicago", Category= "food", *, *)?**
Answer: 3.

Question 3

(40') Given dataset which contains 100 transactions, each line is a transaction and each transaction contains item(s) separated by spaces. Please do frequent pattern mining on this dataset and answer the following questions. If the answer is not an integer, please round the result to 3 decimal places.

1. **Suppose the minimum support is 20. Then:**

   a. **Count the number of frequent patterns.**
   Answer: 30. **Hint:** As long as you understand a particular frequent pattern mining technique, you can find the answers to Question 3 with a simple code. Frequent patterns are those with occurrences greater than or equal to minimal support.

   b. **Count the number of frequent patterns with length 3.**
   Answer: 8. **Hint:** Length 3 means there are exactly 3 items in the pattern. For example, pattern {A,B,C} is of length 3.

c. **Count the number of max patterns.**
Answer: 7. **Hint:** Note that max patterns is neither maximum length nor maximum support. Based on the definition, a pattern **X** is a max-pattern if

(1) **X** is frequent, and

(2) there exists no frequent super-pattern **Y** containing **X**.

Take the example from course slide with transactions

$$\{t_1 : ACD, t_2 : BCE, t_3 : ABCE, t_4 : BE\}$$

with minimum support being 2. The frequent patterns are:

$$\{A, B, C, E, AC, BC, BE, CE, BCE\}.$$

The max patterns are: $\{AC, BCE\}$. Note that you always find frequent patterns first and then eliminate those that have super-patterns (for example BC here).

2. **Suppose we decrease the minimum support to 10.**

a. **Count the number of frequent patterns.**
Answer: 55. **Hint:** For problems (2a.) to (2c.), please refer to problem (1).

b. **Count the number of frequent patterns with length 3.**
Answer: 20.

c. **Count the number of max patterns.**
Answer: 6.

d. **What is the confidence measure of the association rule:** $(C, E) \rightarrow A$**?**
Answer: 0.679. **Hint:** The confidence of $(C, E) \rightarrow A$ is the support of ACE divided by the support of CE.

e. **What is the confidence measure of the association rule:** $(A, B, C) \rightarrow E$**?**
Answer: 0.742.