CS 412: Homework #1

Due on Friday Sept. 30th

Kurt Rudolph
rudolph9

Problem 1

We model the users in a social network as a data cube. Suppose each user has 10 dimensions of information, such as age, gender, city and income. Assume a base cuboid of 10 dimensions contains three base cells: (1) $(b1, b2, a3, a4, a5, \ldots, a9, a10) : count = 10, (2)(b1, a2, b3, a4, a5, \ldots, a9, a10) : count = 20$, and $(3)(a1, b2, b3, a4, a5, \ldots, a9, a10)$:count=50, where $a_i! = b_i, a_i! = a_j$, etc. The count measure of the cube means the number of users who satisfy such information.

(1) How many nonempty cuboids will a full data cube contain?

Solution

(2) How many nonempty aggregate (i.e., non-base) cells will a full cube contain?

Solution

(3) How many nonempty aggregate cells will an iceberg cube contain if the condition of the iceberg cube is "$count \geq 70$"?

Solution

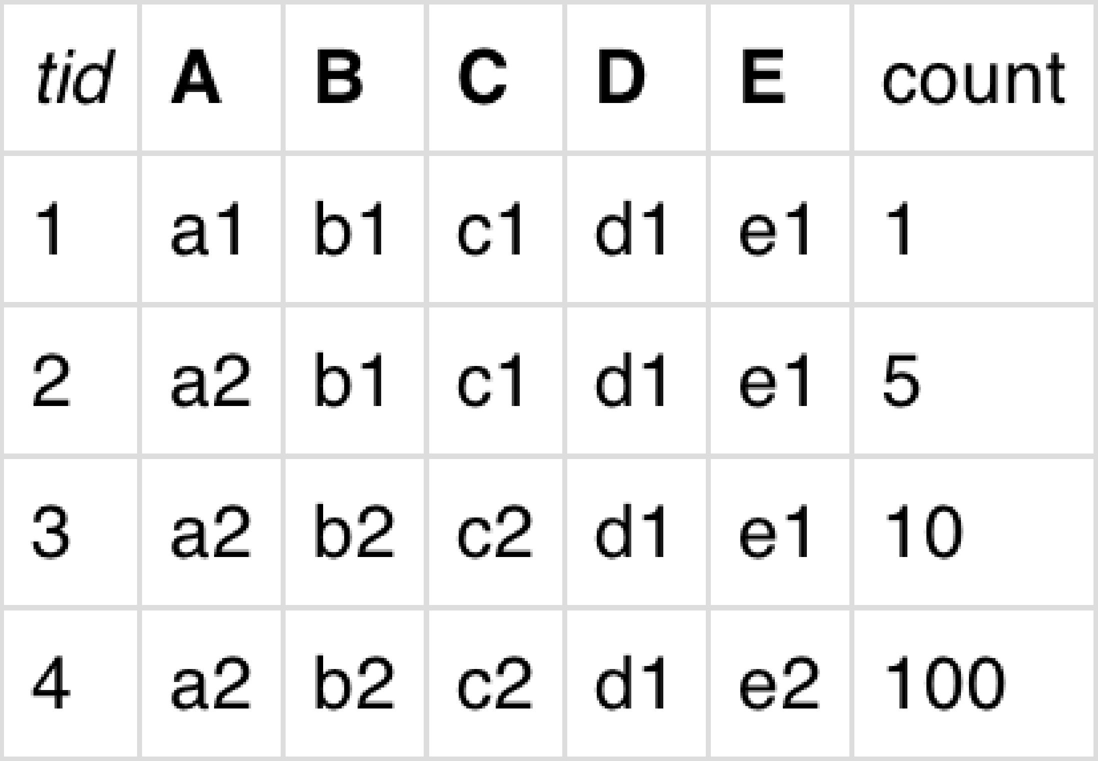
(4) How many closed cells are in the full cube?

Problem 1 continued on next page. . .

Solution

Problem 2

Given the following base cuboid with count as the measure.



tid	**A**	**B**	**C**	**D**	**E**	count
1	a1	b1	c1	d1	e1	1
2	a2	b1	c1	d1	e1	5
3	a2	b2	c2	d1	e1	10
4	a2	b2	c2	d1	e2	100

/2/2

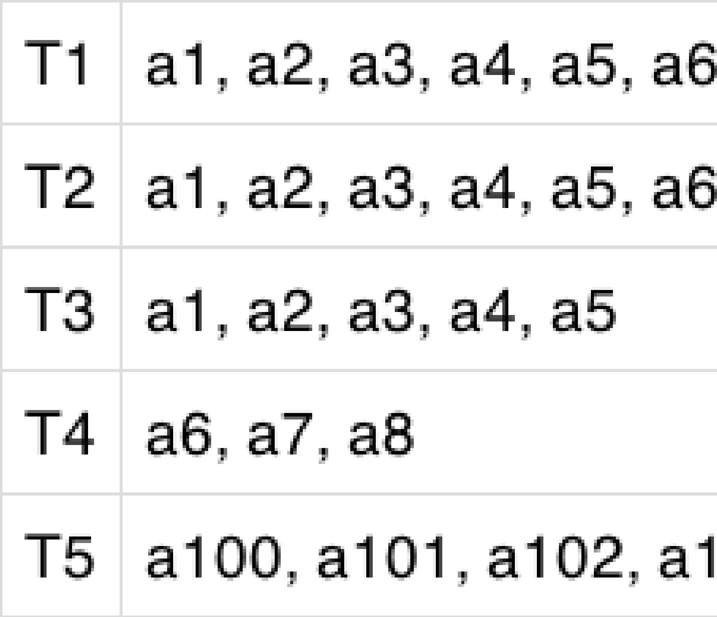
(1) Briefly outline the major steps to compute Shell-Fragment cube (refer to VLDB04 paper High-Dimensional OLAP: A Minimal Cubing Approach), suppose we divide the 5 dimensions into 2 shell fragments: AB and CDE.

Solution

(2) Briefly describe how to compute subcube query (a2,b2,,,? : count())

Solution

Problem 3



T1	a1, a2, a3, a4, a5, a6
T2	a1, a2, a3, a4, a5, a6
T3	a1, a2, a3, a4, a5
T4	a6, a7, a8
T5	a100, a101, a102, a1

Given a database of five transactions ($min_support = 2$): /2/2

(1) How many frequent patterns?

Solution

(2) What is the set of frequent closed patterns (list both pattern and support)?

Solution

(3) What is the set of frequent max-patterns (list both pattern and support)?

Solution

(4) Show an example association rule that matches $(a1, a2, a3, a4, itemX) - > (itemY)[min_support = 2, min_confidence = 70\%]$

Solution

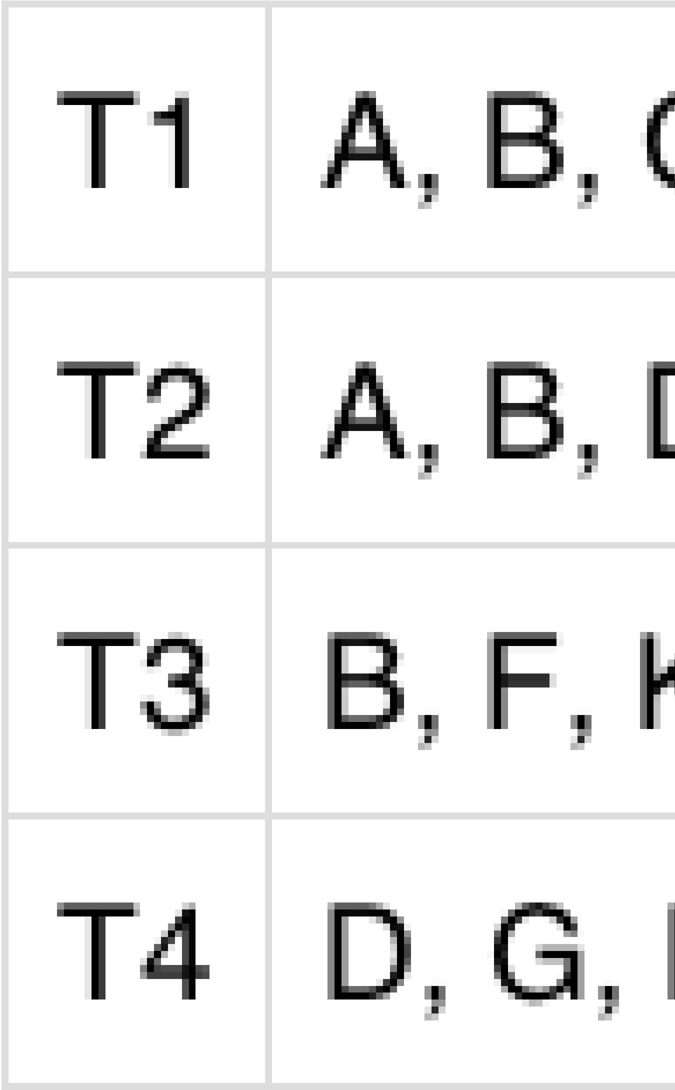
(5) For association rule a1-¿a6, compute the following measures: confidence, lift, kulc.

Solution

(6) Among the above three measures, which ones are null-invariant?

Solution

Problem 4



T1	A, B,
T2	A, B,
T3	B, F,
T4	D, G,

Given a database of four transactions ($min_supprt = 2$): /2/2

(1) Show the major steps to find the frequent patterns using Apriori.

Solution

(2) Show the major steps to find the frequent patterns using FP-Growth (no need to draw the trees).

Solution

(3) Compare the three algorithms: Apriori, FP-growth and ECLAT, by concisely discussing the major differences.

Solution