CS 412 – Introduction to Data Mining

Assignment 2

Problem 1:

(1) Total no. of cuboids in the data cube are $\prod_{i=1}^{d}(L_i + 1) - 1$ excluding the base cuboid if each dimension has L_i levels. Now, since $L_i = 1$, hence we have total $2^d - 1$ cuboids in the given data cube. And each cuboid has n cells except the apex cuboid, we have a total of $n2^d - n + 1$ cells.
Also, we can think like that, each cell generates 2^d cells, so n cells will generate $n2^d$ cells, and the n cells at the apex cuboid are merged into 1 hence we subtract these n-1 cells, hence, the maximum no. of cells in the data cube including the base and aggregate cells are $n2^d - n + 1$.

(2) Considering the case when all the base cells are the same and using the formula for calculating total no. of aggregate cells $\prod_{i=1}^{d}(L_i + 1)$, we get that total no. of cells are 2^d.
Hence, the maximum no. of cells in the data cube including the base and aggregate cells are 2^d

Problem 2:

Since A, B, C and D are binary attributes, let's assume they attain the following values:

$A - \{A_1, A_2\}$, $B - \{B_1, B_2\}$, $C - \{C_1, C_2\}$, $D - \{D_1, D_2\}$

Consider the base case when we are recording (A, *, *, *, M). Now first drilling down on say B would yield (A, B, *, *, M). By symmetry, the observations will hold true for attributes C and D as well.

- Slice on B, Drill on B

1. Consider the situation when we are drilling down and slicing on the same binary attribute (in any order), in which case it is same as simply slicing on the attribute. Since, after drilling down on B, slicing on B_1 would yield (A, B_1, *, *, M) and in reverse order first slicing on B_1 and then drilling down would also yield (A, B_1, *, *, M). Hence, the views are same when we drill down and slice on the same attribute.

2. Now, for other cases when we are drilling down and slicing on different attributes:

- Slice on A, Drill on B

Consider slicing on say A_1 after drilling down on say B (which is (A, B, *, *, M)), it would yield (A_1, B, *, *, M) and drilling down on B after slicing on A_1 (which is (A_1, *, *, *, M)) would be the same (A_1, B, *, *, M).

- Slice on C, Drill on B

Consider slicing on say C_1 after drilling down on say B (which is (A, B, *, *, M)), it would yield (A, B, C_1, *, M) and drilling down on B after slicing on C_1 (which is (A, *, C_1, *, M)) would be the same (A_1, B, C_1, *, M).

- Slice on D, Drill on B

Consider slicing on say D_1 after drilling down on say B (which is (A, B, *, *, M)), it would yield (A, B, D_1, *, M) and drilling down on B after slicing on D_1 (which is (A, *, D_1, *, M)) would be the same (A_1, B, D_1, *, M).

By symmetry, we get the same results by applying drilling on other attributes C, D and slicing along {A, B, C, D} as well, hence we can establish that the results are true in general.

Since, drilling down and slicing are independent operations on a data cube, the two views when we perform slicing and drilling down in any order are same.

Hence, the view V_2 obtained by first drilling down on an attribute from {B, C, D} and then slicing on an attribute from {A, B, C, D} and the view V_3 obtained using the operations in a reverse order will be same.

Experimentally, I used Cubes Viewer to observe this. Below are the plots I got when I used A – Country attribute taking values {Europe, Africa}, B – Product Category taking values {Books, Sports}, C – Year attribute taking values {2014, 2015}, D.

It can be observed that same charts were obtained on performing the two operations in any order.

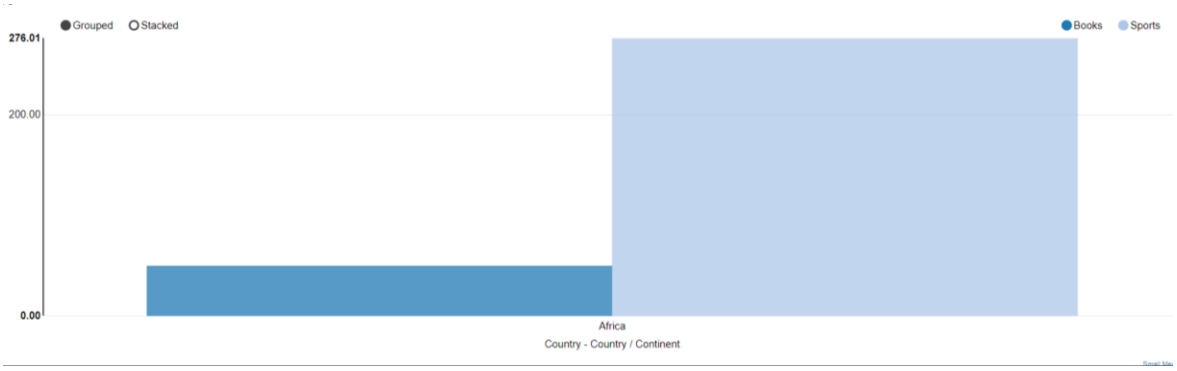

Fig 1: Chart obtained by Drilling down on B and slicing on A (Africa) or vice versa.

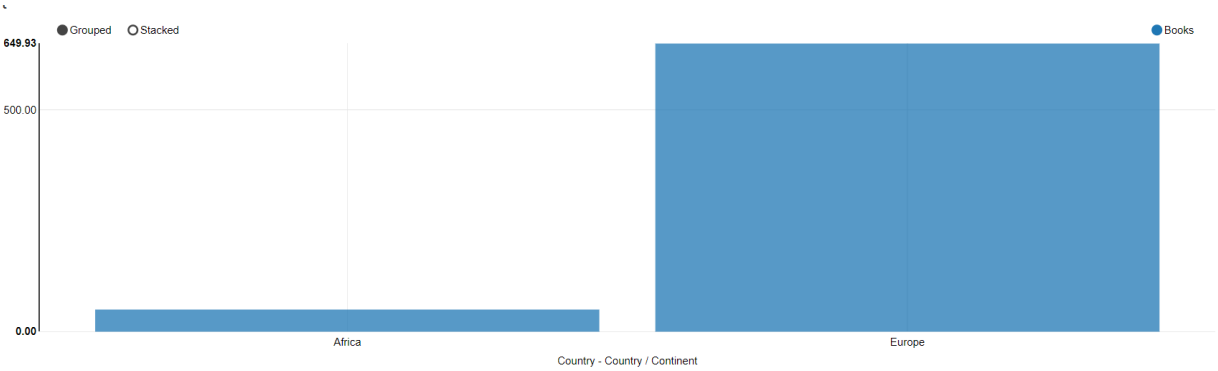

Fig 2: Chart obtained by Drilling down on B and slicing on B (Books) or vice versa.

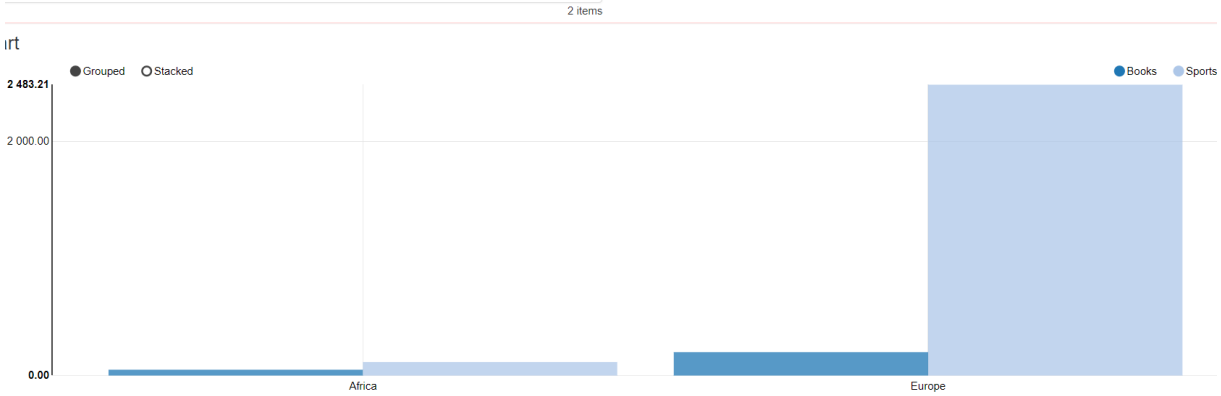

Fig 3: Chart obtained by Drilling down on B and slicing on C (Year) or vice versa.

Hence, verifying experimentally the claim that V_2 and V_3 will have the same view.

Submitted by:

Rachneet Kaur, Net ID: rk4