Databases and Data Mining

Hints for Databases Assignment 2

23.10.2018

The following more detailed explanation of one of the examples used during the lecture, on the slides and in the book, should help you to get on track with in particular (but not necessarily only) Question 3 of Assignment 2.

Suppose that there are only two base cells, {(a1, a2, a3, . . . , a100), (a1, a2, b3, . . . , b100)}, in a 100-dimensional base cuboid.

- The number of non-empty (having count >= 1) aggregate cells is 2^101 – 6. Why?
  (In fact, there is a typo on the slides: 2^101 – 4 is the number of all non-empty cells, i.e., including both base and aggregate cells; the respective example in the book is correct.)
  Well, each base cell generates 2^100 − 1 aggregate cells, that is one cell in each of the 2^100 cubiods in the complete cube, minus 1 that is the base cell (in the base cubiod) itself. Thus, the two base cells generate 2×(2^100 − 1) = 2^101 − 2 aggregate cells, however, four of these cells are thus counted twice as they overlap because the first two dimension have only 1 distinct value each (a1 and a2, respectively) in the base cubiod. These four cells are, hence: (a1, a2, ∗, . . . , ∗), (a1, ∗, . . . , ∗), (∗, a2, ∗, . . . , ∗), and (∗, ∗, . . . , ∗). Therefore, the total number of aggregate cells generated is 2^101 − 6, while the total number of cells in the complete cube is 2^101 – 4.
- The number of non-empty aggregate cells in the iceberg cube with condition having count >= 2 is 4. Which and why?
  There are only 4 aggregate cells that aggregate more than one base cell (and hence have count > 1). These are the same 4 aggregate cells we subtracted above: {(a1, a2, ∗, . . . , ∗), (a1, ∗, ∗, . . . , ∗), (∗, a2, ∗, . . . , ∗), (∗, ∗, ∗, . . . , ∗)}.
- The closed cube consists of only 3 closed cells. Which and why?
- There are only two different measure values in this example: the four aggregate cells mentioned above have measure (count) 2, while all other aggregate cell and the two base cells have measure (count) 1. Of the above four aggregate cells with measure 2, the 0-dim. Apex {(∗, ∗, ∗, . . . , ∗)} and the two 1-dim. aggregates {(a1, ∗, ∗, . . . , ∗)} are ancestors (generalizations) of the 2-dim. aggregate {(a1, a2, ∗, . . . , ∗)}, (i.e., the latter is a descendant (specialization) of the former where one (or more) *-value(s) is/are replaced by non-* values), and thus not closed. Likewise, all aggregates with measure 1 are ancestors (generalizations) of one of the two base cells. Hence, the closed cube consists of the following 3 closed cells: {(a1, a2, a3, . . . , a100), (a1, a2, b3, . . . , b100), (a1, a2, ∗, . . . , ∗)}.

(Correct) adaptation/“extrapolation” from 2 to 3 base cells and from 100 to 9 dimensions should yield the answer(s) to Question 3.

And here are two 4-d examples with two base cells:

4-d base cells: 2 (a1,a2,a3,a4) : 1 (a1,a2,b3,b4) : 1 3-d aggr. cells: 8 (*,a2,a3,a4) : 1 (*,a2,b3,b4) : 1 (a1,*,a3,a4) : 1 (a1,*,b3,b4) : 1 (a1,a2,*,a4) : 1 (a1,a2,*,b4) : 1 (a1,a2,a3,*) : 1 (a1,a2,b3,*) : 1 2-d aggr. cells: 11 (*,*,a3,a4) : 1 (*,*,b3,b4) : 1 (*,a2,*,a4) : 1 (*,a2,*,b4) : 1 (*,a2,a3,*) : 1 (*,a2,b3,*) : 1 (a1,*,*,a4) : 1 (a1,*,*,b4) : 1 (a1,*,a3,*) : 1 (a1,*,b3,*) : 1 (a1,a2,*,*) : 2 1-d aggr. cells: 6 (*,*,*,a4): 1 (*,*,*,b4): 1 (*,*,a3,*): 1 (*,*,b3,*): 1 (*,a2,*,*): 2 (a1,*,*,*): 2 0-d aggr. cell: 1 (*,*,*,*): 2 total: non-empty base cells: 2 non-empty aggr. cells: 2 * (2^4 – 1) – 4 = 2^5 – 2 – 4 = 2^5 – 6 = 26 non-empty cells: 2 + 2^5 – 6 = 2^5 – 4 = 28	4-d base cells: 2 (a1,c2,c3,c4) : 1 (c1,b2,c3,c4) : 1 3-d aggr. cells: 8 (*,c2,c3,c4) : 1 (*,b2,c3,c4) : 1 (a1,*,c3,c4) : 1 (c1,*,c3,c4) : 1 (a1,c2,*,c4) : 1 (c1,b2,*,c4) : 1 (a1,c2,c3,*) : 1 (c1,b2,c3,*) : 1 2-d aggr. cells: 11 (*,*,c3,c4) : 2 (*,c2,*,c4) : 1 (*,b2,*,c4) : 1 (*,c2,c3,*) : 1 (*,b2,c3,*) : 1 (a1,*,*,c4) : 1 (c1,*,*,c4) : 1 (a1,*,c3,*) : 1 (c1,*,c3,*) : 1 (a1,c2,*,*) : 1 (c1,b2,*,*) : 1 1-d aggr. cells: 6 (*,*,*,c4): 2 (*,*,c3,*): 2 (*,c2,*,*): 1 (*,b2,*,*): 1 (a1,*,*,*): 1 (c1,*,*,*): 1 0-d aggr. cell: 1 (*,*,*,*): 2 total: non-empty base cells: 2 non-empty aggr. cells: 2 * (2^4 – 1) – 4 = 2^5 – 2 – 4 = 2^5 – 6 = 26 non-empty cells: 2 + 2^5 – 6 = 2^5 – 4 = 28