Hasso-Plattner-Institute
Knowledge Discovery and Data Mining
Prof. Emmanuel Müller

November 4th, 2015

Exercise 1 for the lecture
Big Data Analytics
WS 2015/2016

Please prepare solutions for the tutorial November 11th in order to participate actively during the tutorial session. Solutions will be presented by students and discussed during the tutorial.

Exercise 1) KDD process

Call the steps of the KDD process and their primary objectives.

Exercise 2) Incremental Aggregation

Are the following measures algebraic, holistic or distributive? Provide a detailed reasoning (computation rule or formal proof) for your answer.

1. average value
2. standard deviation
3. mode

Exercise 3) Bayes

Prove the following equation with the help of the definition of conditional probabilities:

$$p(o) * p(c_j \mid o) = p(o \mid c_j) * p(c_j)$$

Exercise 4) OLAP Operations (SQL):

Given the following example database of product sales consisting of

i. a fact table for product sales (*transactionID* is the key for this table) containing information about the date, the sold product, the city and the number of sold products,
ii. a dimension table with information about cities and their corresponding state and country (*city* is the key for this table),
iii. a dimension table with information about products (*productID* is the key for this table) and
iv. a dimension table with information about dates (this would not be done like this in the "real world", but we nevertheless introduce this table…).

location : Tabelle

city	state	country
Los Angeles	California	USA
Mexico City	Mexico	Mexico
Miami	Florida	USA
Nelson	British Columbia	Canada
Orlando	Florida	USA
San Francisco	California	USA
Toronto	Ontario	Canada
Vancouver	British Columbia	Canada
Waterloo	Ontario	Canada

product : Tabelle

productID	productName
P1	product 1
P10	product 10
P2	product 2
P3	product 3
P4	product 4
P5	product 5
P6	product 6
P7	product 7
P8	product 8
P9	product 9

date : Tabelle

day	month	quarter	year
30.11.2004	112004	042004	2004
01.01.2005	012005	012005	2005
05.01.2005	012005	012005	2005
15.04.2005	042005	012005	2005
18.04.2005	042005	022005	2005
20.04.2005	042005	022005	2005
21.04.2005	042005	012005	2005

sales : Tabelle

transactionID	productID	day	city	no_of_products
1	P10	21.04.2005	Los Angeles	5
2	P05	01.01.2005	Miami	4
3	P05	30.11.2004	Orlando	2
4	P1	15.04.2005	Toronto	2
5	P3	05.01.2005	Toronto	4
6	P5	30.11.2004	Miami	7
7	P10	18.04.2005	Miami	5
8	P3	01.01.2005	Miami	4
9	P1	30.11.2004	Orlando	2
10	P3	20.04.2005	Los Angeles	5
11	P10	01.01.2005	Miami	4
12	P1	30.11.2004	Orlando	2

a) Draw the star schema for this database.

b) Write the following queries in SQL:

   a. show the aggregated product sales for each product, country and day
   b. show the aggregated product sales for each product, country and month
   c. show the aggregated product sales for each product, country and quarter
   d. show the aggregated product sales for each product, country and year
   e. show the aggregated product sales for each product, state and day
   f. show the aggregated product sales for each product, city and day
   g. show the product sales for product P5 and P10
   h. show the product sales for Canada
   i. show the product sales for product P10 and Miami.

c) Specify the OLAP operations of part b).

Exercise 5) R-Trees:

a) How large is an R-Tree that stores 1.000, 100.000, 1.000.000, 10.000.000, 1.000.000.000 objects. Assume parameters M = 400 and m = 200 for your calculations. Derive the results by calculating the MBRs that can be stored on each level of the tree.

b) Provide the order of nodes that need to be checked for a 2-NN query on the following R-Tree (as given in the lecture). Please discuss possible alternative options in processing the MBRs during the query processing. Provide two alternative lists of accessed MBRs, compare these two alternatives, and discuss how one could guide the query processing during search.