Data Mining
Homework Assignment #12

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

May 15, 2014

Task 1

State why for the integration of multiple heterogeneous information sources many companies in industry prefer the update-driven approach (which constructs and uses data warehouses), rather than the query-driven approach (which applies wrappers and integrators). Describe situations where the query-driven approach is preferable over the update-driven approach.

Task 2

A data warehouse can be modeled by either a star schema or a snowflake schema. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. Give your opinion of which might be more useful on practice and state the reasons behind your answer.

Task 3

Explain the terms slice and dice, drill-down and roll-up. Provide examples and/or illustrations

Task 4

Suppose that a data warehouse consists of the four dimensions: date, visitor, location, and movie, and the two measures, count and charge, where charge is the fare that a visitor pays when watching a movie on a given date. Visitors may be students, adults, or seniors, with each category having its own charge rate.

a. Draw a star schema diagram for the data warehouse.

b. Starting with the base cuboid [date, visitor, location, movie], what specific OLAP operations should one perform in order to list the total charge paid by student visitors at Cinamon Cinema in 2004?

Task 5

Rewrite the following query by replacing GROUP BY CUBE(...) clause into an equivalent query using UNION and simple GROUP BY:

```
SELECT product, year, city, sum(price*vol)
FROM Orders
GROUP BY CUBE(product, year, city);
```

Task 6

Use data sample from US census 2000 `http://biit.cs.ut.ee/~vilo/edu/Data/census2000/extract_medium.csv.gz`. Characterize the relationship between salary and income in relation to State and Age of a person. (hint: use pivot tables of Excel/OO/LibreOffice). If you use Excel, add heatmap on top of the pivot table. To illustrate some of the found relationships, add a screenshot.

Inspired and compiled from Data Mining: Concepts and Techniques by Han, J., Kamber, M.