Hand-in 4: OLAP and data mining

> *This hand-in must be handed in as a single PDF file by each group using the LearnIT system no later than*
>
> ***Friday November 30, 23.59*.**
>
> *Please include the group number in the file name.*

In this hand-in you will be doing analytics on a data set of movie ratings from MovieLens (movielens.umn.edu). It contains about 1 million user ratings of movies (on the scale 1-5), each associated with a unique user ID and the date of the rating. For each user, gender, approximate age, occupation, and zip code is recorded. Finally, information about the genre(s) of each movie is available. The main data is available as three relations:

```
    user(id,gender,age,occupation,zip)
    rating(userId,movieId,rating,time)
    movieGenre(genreId,movieId)
```

These relations use numerical codes for occupation, movie ID, and genres. Textual descriptions of these codes are available as:

```
    occupation(id,description)
    movie(id,title)
    genre(id,name)
```

Finally, a separate relation with information about zip codes is available:

```
    zipcode(zip,city,state,lattitude,longitude,timezone,dst)
```

In the LearnIT directory "data" you can download a (zip'ed) file `movielens.sql` containing this data, prepared for importing into MySQL. (There is also a bigger file, `movielensXL.sql`, if you feel like trying some of the tasks with a bigger, but not as detailed, data set.)

Tasks

Your first task is to perform *data cleaning*. In particular, we wish to remove all data that lacks proper foreign key references. For example, there are tuples in `user` that do not refer to a tuple in `zipcode`. Another issue is that age approximations can be made more accurate. In the data set, age 1 means "Under 18", age 18 means "18-24", age 25 means "25-34", etc., up to 56 which means "56+". Modify the ages so that they are a "best guess" for the range; clearly, this is not an exact science.

Second, you should choose a way of *enriching* the data set. For example, you could make a table with the median or average household income for each zip code. This data can be downloaded in XLS format from http://www.psc.isr.umich.edu/dis/census/Features/tract2zip/index.html. To get the data into MySQL, export as a comma-separated (CSV) file, and use MySQL's `LOAD DATA LOCAL INFILE` syntax to load it into a table. Alternatively, integrate with information in the IMDB database you worked with previously. (NB! Movie IDs and titles are not identical.)

The third task is to create a relational OLAP model for the data, according to the principles discussed in the class. Since MySQL does not support materialized views, you will need to construct separate tables with pre-aggregated data, and put indexes on these. B-tree indexes will suffice, even though they may be slower than bitmap indexes in this context. Motivate your choice of pre-aggregation with a few usage scenarios, and give corresponding example SQL queries (probably working on the pre-aggregated tables rather than the whole data set).

Finally, after the data mining lecture you should think about what kinds of data mining tasks would be relevant and interesting on the extended MovieLens data set.

To be handed in

- Name of all group members who contributed to the hand-in.
- A short description of the data cleaning and enrichment performed, including the MySQL commands used.
- A description of the relational OLAP model created (using the terms *fact, measure, dimension*), along with the DDL.
- A description of the pre-aggregation tables and indexes created, incl. DDL.
- For each group member, an *interesting fact* about the ratings data, found using OLAP exploration. E.g., "those who liked *Bowling for Columbine* live in zip codes with average household income that is…". There will be a **prize** for the best fact found, as judged by the TAs.
- A description at least two data mining algorithms that could be run on the data set. For each, state what the input would be, and what kind of result you would expect the data mining algorithm to return.

Course goals covered by this hand-in

After the course the students should be able to:

- suggest a conceptual and physical design of an OLAP system.
- suggest an abstract model suitable for a given data mining task.