Hand-in 2: SQL queries and indexing

This hand-in must be handed in as a single PDF file by each group using the LearnIT system no later than

Friday October 12, 23.59.

You previously defined a relational schema for your movie database based on IMDB.com. Based on the feedback on the first hand-in, you may want to update your E-R model. Also, you now get additional information on what data must be stored in the database. In particular, the database schema should allow the database to contain enough information to answer the following queries:

1. How many Danish language movies are in the database?
2. For each year, what is the total number of votes to movies from that year?
3. Which movies have John Travolta and Uma Thurman starred in together?
4. How many actors and directors have a first name starting with Q?
5. How many students from the class rated at least 10 movies?
6. What is the name and birth year for all actors in Pulp Fiction? Your query should list the actors in increasing order of birth year.
7. What are the titles and years of all movies from the 1980s where John Travolta starred?
8. What are the top-5 highest rated movies from the 1990s according to IMDB users?
9. What are the top-5 highest rated (average) from the 1990s according to at least 4 students from the class?
10. In 1994, what was the average IMDB rating of a movie for each language?
11. Which actors in *Pulp Fiction* have never, before or after, starred in the same movie as one of the other actors in *Pulp Fiction*?
12. Which movie starring John Travolta has the highest IMDB user rating?
13. How many actresses have *not* been alive at the same time as Charles Chaplin?
14. What is the average rating of movies from each genre?
15. For each genre, what is the number of student ratings of a movie from that genre? List only genres with at least 10 ratings.
16. Which movie has the largest number of 2-link references? (If A refers to B, and B refers to C, then we say that A has a 2-link reference, through B, to C. If there are several paths leading from A to C, we count all of them.)
17. How many actors have also been active as director of at least one movie?
18. Which two genres are most often linked to the same movie? (Note that each movie has a set of genres.)

You should implement and test the SQL queries in MySQL on real IMDB data. The data should be loaded into a separate database (see instructions below) and then transferred to your own schema using SQL.

Finally, for each query that does not have satisfactory performance you should define one or more indexes that improve the running time. You should use `EXPLAIN` in MySQL to see that your indexes are used (or at least possibly used) for the queries.

To be handed in

- Name of all group members who contributed to the hand-in.
- A graphical presentation of the (possibly revised) E-R model. Comment on changes made since hand-in 1.
- An updated schema, written in SQL, corresponding to the E-R model. Also hand in the SQL commands used to transfer data to your schema.
- SQL for at least 15 of the queries described above. Unless a query is self-explanatory, give a short explanation of its correctness.
- A list of the indexes created (not including indexes automatically built on primary keys). Report timings of all your SQL queries before and after index creation. (You may have to run each query twice to ensure that the relevant parts of the index and relation are in RAM – use the time for the second run.)
- A transcript of a database terminal session running all queries in MySQL on the loaded data set, and also showing query plans (using EXPLAIN). Discuss how the observed behavior fits with your understanding of indexes.

Course goals covered by this hand-in

After the course the students should be able to:

- Write SQL queries, involving multiple relations, compound conditions, grouping, aggregation, and subqueries.
- Decide if a given index is likely to improve performance for a given query.

Instructions for loading IMDB data

In the LearnIT directory "data" you can download a (zip'ed) files `imdb.sql` and `ratings.sql` containing the IMDB data and class ratings in a fixed schema. Observe that this data is only provided for use within the course; see files found at www.imdb.com/interfaces for terms of use. To load the data first create a separate database in MySQL using `CREATE DATABASE <dbname>` on your own machine, or using database self service if you are using ITUs MySQL server. Unzip the files, open the command line window (or ssh to `ssh.itu.dk`), and type something like the below, with `<dbname>` replaced by the name of your database, `localhost` replaced by `mysql.itu.dk` if you are using the ITU server, and `root` replaced by your database user name (if you created one other than root):

```
mysql -u root -p -h localhost <dbname> < imdb.sql
```

If you did not put the `mysql` command line tool in your path, you will have to include its location in the file system. For example, in my installation on OSX it is `/usr/local/mysql/bin/mysql`, and on Windows it is typically in `C:\"Program Files"\MySQL\"MySQL Server 5.5"\bin\mysql.exe`
If you need help loading the data, please contact your teaching assistant.