Introduction to Database Design, E2015
IT University of Copenhagen
Carsten Schürmann

Assignment 2

Theoretical Part

Problem 1: Relational Algebra [21 points]

Consider the relations of a Jobs database. These following three relations record people, their skills and their endorsements.

- `Member(userID,name)` For each member we record the userID and name.
- `Skill(skillID,skillName)` For each skill we record the skillID and the skillName.
- `Endorsement(userID,skillID,prof,n)` Each row shows which person has what skill, at what proficiency level prof (eg., months of experience), and the count n of endorsements from other members.

and sample instance:

`Member`

userID	name
U001	Jesper
U002	Jette
U003	Josef
U004	Jane
U005	Jens
U006	Jan

`Skill`

skillID	skillName
S1	SQL
S2	Java
S3	Ruby
S4	SAP
S5	R

`Endorsement`

userID	skillID	prof	n
U003	S1	18	5
U003	S2	46	5
U003	S4	17	5
U002	S4	10	4
U001	S4	15	1
U006	S4	21	1
U005	S4	31	3
U004	S5	45	4

1. [3 point] Which of the following is the meaning of the expression $\sigma_{n<5}$ (`Endorsement`)

   (a) It lists all the `n` values, that are less than 5, eliminating duplicates (i.e., $\{1, 3, 4\}$ in our case).

   (b) It lists all Endorsement tuples (`userID`, `skillID`, `prof`, and `n`) with less than 5 endorsements.

   (c) It lists the `n` value for each `Endorsement` tuple, and it rounds it down to 5, if higher than 5.

   (d) None of the above. The real answer is ....

2. [3 point] We want to list the mature skills, that is, the `skillNames`, for which there is at least one veteran (defined as `prof`$>$ 36 months of experience). Which, if any, of the following expressions achieves that?

   (a) $\sigma_{\texttt{skillName}}(\pi_{\texttt{prof}>36}(\texttt{Skill} \bowtie \texttt{Endorsement}))$

   (b) $\pi_{\texttt{skillName}}(\sigma_{\texttt{prof}>36}(\texttt{Skill} \bowtie \texttt{Endorsement}))$

   (c) $\pi_{\texttt{skillName}}((\sigma_{\texttt{prof}>36}(\texttt{Skill})) \bowtie \texttt{Endorsement})$

   (d) $\sigma_{\texttt{skillName}>36}(\pi_{\texttt{prof}}(\texttt{Skill} \bowtie \texttt{Endorsement}))$

   (e) None of the above. The real answer is ....

3. [5 points] For the following expression:

   $$\sigma_{\texttt{n}<4}(\texttt{Member} \bowtie \texttt{Endorsement})$$

   (a) Describe in English what the expression does.

   (b) How many, and which are the columns (= attributes) in the answer?

   (c) How many tuples are in the answer?

   (d) List all the tuples in the answer, as a table.

4. [5 points] For the following expression:

   $$\pi_{\texttt{userID,skillID}}(\texttt{Endorsement}) \div \pi_{\texttt{skillID}}(\sigma_{\texttt{userID=}"U005"}(\texttt{Endorsement}))$$

   (a) Describe in English what the expression does.

   (b) How many, and which are the columns (= attributes) in the answer?

   (c) How many tuples are in the answer?

   (d) List all the tuples in the answer, as a table.

5. [5 points] For the following expression:

   $$\pi_{\texttt{E.userID,E1.userID}}(\varrho_{\texttt{E}}(\texttt{Endorsement}) \bowtie_{\texttt{E.skillID=E1.skillID}\wedge\texttt{E.userID>E1.userID}} \varrho_{\texttt{E1}}(\texttt{Endorsements}))$$

   (a) Describe in English what the expression does.

   (b) How many, and which are the columns (= attributes) in the answer?

   (c) How many tuples are in the answer?

   (d) List all the tuples in the answer, as a table.

Problem 2: Relational Tuple Calculus [15 points]

We continue with the Job database from above:

1. [5 points] For the following expression

   $$\{t | \exists e \in \texttt{Endorsement}.e.\texttt{skillID} = ``S1'' \wedge e.\texttt{userID} = t.\texttt{userID}\}$$

(a) Describe in English what the expression does.

(b) How many, and which are the columns (= attributes) in the answer?

(c) How many tuples are in the answer?

(d) List all the tuples in the answer, as a table.

2. [5 points] For the following expression

$$\begin{aligned}\{t|\ & \exists e_1 \in \texttt{Endorsement}.\exists e_2 \in \texttt{Endorsement}.\\ & e_1.\texttt{skillID} = e_2.\texttt{skillID}\\ & \wedge e_1.\texttt{userID} > e_2.\texttt{userID}\\ & \wedge t.\texttt{user1} > e_1.\texttt{userID}\\ & \wedge t.\texttt{user2} > e_2.\texttt{userID}\}\end{aligned}$$

(a) Describe in English what the expression does.

(b) How many, and which are the columns (= attributes) in the answer?

(c) How many tuples are in the answer?

(d) List all the tuples in the answer, as a table.

3. [5 points] For the following expression

$$\begin{aligned}\{t|\ & \exists e \in \texttt{Endorsement}.\exists m \in \texttt{Member}.\\ & e.\texttt{userID} = m.\texttt{userID}\\ & \wedge t.\texttt{name} = m.\texttt{name}\\ & \wedge e.\texttt{skillID} = "S1"\\ & \wedge e.\texttt{prof} > 20\}\end{aligned}$$

(a) Describe in English what the expression does.

(b) How many, and which are the columns (= attributes) in the answer?

(c) How many tuples are in the answer?

(d) List all the tuples in the answer, as a table.

Practical Part

Problem 3: SQL Queries [51 points]

Using the imdb database schema that you can find on learn-it, please write SQL queries for the following questions.

1. How many Danish language movies are in the database?

2. For each year, what is the total number of votes casted for all movies produced that year?

3. How many actors and directors have a first name starting with C?

4. What are the names and birth years for all actors in Pulp Fiction? Your query should list the actors in increasing order of birth year.

5. Which movies have either John Travolta or Uma Thurman, but not both starred in?
6. Which actors in Pulp Fiction have never, before or after, starred in the same movie as one of the other actors in Pulp Fiction?
7. What are the titles and years of all movies since 1980 where John Travolta costarred with Samuel L Jackson?
8. What are the top-5 highest rated movies from the 1990s according to IMDB users?
9. In 1994, what was the average IMDB rating of a movie for each language?
10. Which movie starring John Travolta has the highest IMDB user rating?
11. How many actresses did not live at the same time as Charles Chaplin?
12. Who was the shortest actor to star in an action movie?
13. What is the average rating of movies from each genre?
14. For each genre, what is the number of student ratings of a movie from that genre? List only genres with at least 10 ratings.
15. 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.)
16. How many actors have also been active as director of at least one movie?
17. 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. Submit all of your queries, together with a detailed transcript!

Problem 4: Indexing [13 points]

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.

Please submit the 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.)

Instructions for loading IMDB data

From the course homepage you can download a (ziped) 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.

The second hand-in should be handed in by each group no later than

Friday 9.10.2015, 23:55

It suffices that one group member sends the solution as a single PDF file with a file name that includes the group number. Late hand-ins will not be corrected. Please upload your reply to learn-it. This assignment will be graded and contributes with 15% to your final grade.