CSCI620 Phase 1 : IMDB Dataset

Ketan Kokane
Rochester Institute of Technology
1 Lomb Memorial Dr,
Rochester, NY 14623
kk7471@rit.edu

Siddarth Sargunaraj
Rochester Institute of Technology
1 Lomb Memorial Dr,
Rochester, NY 14623
sxs2469@rit.edu

Ameya Nagnur
Rochester Institute of Technology
1 Lomb Memorial Dr,
Rochester, NY 14623
an4920@rit.edu

Kavya Kotian
Rochester Institute of Technology
1 Lomb Memorial Dr,
Rochester, NY 14623
kk2014@rit.edu

ABSTRACT

This *project* analyses the dataset used by IMDb. There is an ER diagram representing the entities constituting the dataset and relationships between them. Based on this ER diagram, we build a Relational Schema that helps us write queries to create tables for this database. We improve the database with indexing and create queries to provide efficient access of data to the users..

Keywords

Data Management, ER Diagram, Relational Scheme

1. WHAT HAS BEEN DONE?

- Analysis of the entities in the IMDB dataset
- Convert the dataset entities to an ER diagram
- Improve upon the ER Diagram
- Create a Relational Schema

2. FUTURE PLANS

We will normalize the database to get it into the BCNF form to enable it's population. Then we have to populate the database. For this, we will get the tsv files for the dataset provided by IMDb. We use a python script to read these tsv files and load the database with tuples for every table. Now, we add indexes to the tables for making the database more efficient for access with queries. We create single-table, multi-table as well as nested queries to make database access easy for user. Queries like Retrieve all movies with the genre as comedy, Retrieve all movies by a certain director with rating > 8.0 etc. Finally, we will update the indexing of the tables to enhance the working of these queries.

3. ANALYSIS OF THE ENTITIES IN THE IMDB DATASET

Only entities defined in the dataset are movies(TV series can also be considered as movies) and cast.

IMDb provides data about movies and TV Series, which will be referenced as titles further on.

Every title has a tconst which is unique for every movie and is shared by the series episodes, Has title type (movie, short, tvseries, tvepisode, tvmovie, video, tvMiniSeries, tvShort),ordering, Primary title, Original title, Is the movie adult, Release year of the title, End year, Run time of the title, Genres the title belongs to which is maximum of 3 different genres per title, language the movie was made in, type of title (categories set by IMdb itself), additional data about the movie (can be in text format), every movie has a rating associated with it which is average rating provided by the users. Every movie has principal cast

Imdb provides below data about every cast members of every title

Every cast mumber has nconst which is unique primary name, birth year, death year, primary profession (can have up to 3 different professions). Popular titles the person are known for Character the person plays in the given tconst

4. RELATIONAL SCHEMA

Title (tconst, ordering, primary title, imdb title type, is adult ?, average rating, original title, runtime minute, start year, regional title, end year, regional title description)
PK: tconst, ordering

Title_genres (tconst,ordering, genres)
PK:(tconst, ordering, genres)
FK: tconst, ordering

person(nconst, primary name, birth year, death year)
PK: nconst

Person primary profession(nconst, primary profession)
PK:nconst, primary profession

Person know for title(nconst, tconst,ordering)
FK1:nconst **FK2**: tconst, ordering

Title pricipal cast character (tconsts, ordering, nconst, character) **FK**: tconsts, ordering, nconst

Episode name(parent tconts, tconts,ordering)
FK: parent tconts, tconts

5. DELIVERABLES

- Python scripts to load data into SQL using tsv files.
- ER diagram.
- Relational schema of the designed database
- Normalizing the database to 2NF form.
- SQL script to create database tables.