CSCI-620 Data Management with the IMDb Dataset

[Understanding, modeling, and developing tools to interact with the IMDb dataset]

Aishwarya Rao
ar2711@rit.edu

Apurav Khare
ak2816@rit.edu

Martin Qian
jq3513@rit.edu

Prateek Kalasannavar
pk6685@rit.edu

ABSTRACT

This project aims to explore a dataset by understanding it, modeling it to a normalized relational schema so that it can be stored and retrieved from a relational database management system. The project also focuses on developing an interface that allows fast and easy access to the dataset by abstracting complex query scenarios, like search by specific parameters within and across tables, and aggregate queries.

1. THE IMDB DATASET

The dataset chosen for this project is the IMDb dataset, which is a vast collection of movie, series, actor, and rating information. The data is chosen as it is publicly available and regularly updated. Though the data contains logically separated files for the basic metadata, alternate titles, episodes, cast, and rating, it does not adhere to ACID properties. For instance, the data in the column "knownForTitles" contains an array. There are also certain columns, like "job" and "characters" that contain missing values represented as "\N". The dataset also contains columns that do not have any data at all, like "attributes". These features pose a challenge in modeling and creating a robust query system that can access the data efficiently. It is considerably large and requires critical modeling and design to access it efficiently through queries.

1.1 Dataset Features

The IMDb dataset consists of seven zipped files including the title, crew, actor names, and ratings. Each of these files have their own attributes represented in the tab spaced formatting.

The core entity set of the data is the title, which represents a movie or a series or a short. It has basic properties like title, start/end year, and genres (array). A title may have alternative titles, which are used when a title is released on a different media format or in another language or region, for instance, the US title "False Colors" is called "False Colours" in the British release. Titles that are released serially are also associated with episodes and seasons. It is interesting to note that season and episode together identify a episode (season 1 has episode 1, and season 2 has episode 1). A surrogate key exists in this file which is unique per season per episode. The crew of the title contains the writers and directors for the title. The titles can also have a rating. This dataset provides the average rating and the number of votes that led to the average; fine-grained information about each review and its rating is not available in the dataset.

The other core entity set of the data is the "name", which refers to the metadata of a person that has participated in a title. This relation is established by the "title.principals" file which maps titles to names.

The challenges offered by this dataset include modeling of the data by keeping in mind both the normalization constraints as well as the current database structure, handling multi-value attributes that are present in some of the tables, and designing a script and model capable of handling missing values for some attributes.

2. PROJECT OUTLINE

2.1 Data Modeling

In this phase, we understand and explore the data thoroughly to identify entities, their relations with each other, and all relevant constraints. The outcome of this phase will be an ER diagram that summarizes the observations by week 4. This phase would be led by Aishwarya.

2.2 Schema Identification

This phase involves the identification and mapping of the relational schema from the entity- relationship Model, as well as choose appropriate constraints and keys required to maintain data integrity and avoid redundancies. Prateek will be leading this phase. It ends by week 5 with the predicted deliverable as the populated relational database system.

2.3 Create Queries

The next phase in the project will involve identifying possible use-case scenarios of the dataset and building equivalent queries to handle these requests. These queries will include searching by title, actor, genre, or other attributes. It will also allow sorting by year, title, or any other appropriate parameter. This phase will also require testing and refining the queries that are built.

Some possible query scenarios that would aid the user to better access the data are:

- Case-insensitive fuzzy search across tables and columns. For example, finding title across the original and alternate titles, and by principals, genres, or crew.
- Windowed queries, for instance, ranking the titles by rating across release dates.
- Pivoted queries to find a title across release media.
- Aggregated summaries of actors and their contributions to titles.

Martin will be leading this phase. The deliverable for this phase is the SQL script for database access and is expected to be completed by week 7.

2.4 User Interface Development

In this phase, we develop an user-friendly interface for potential customers to interact with the database without having to create their own queries. The focus of this phase will be on ease of access, understandable user interface with popular web technologies such as HTML and javascript. The user interface is then linked to the database and tested rigorously. This phase would be led by Apurav and is expected to be completed by week 9.