Anime Recommendation System

CSPB4502 Spring 2022

Matthew Fournier
Matthew.Fournier@colorado.edu

Christie Hui
chhu9206@colorado.edu

PROBLEM STATEMENT

The aim of this project is to create an anime recommendation system for users based on existing user data obtained from MyAnimeList.net, which is essentially a cataloging site which allows users to keep lists of which anime they've seen/plan to see, view user reviews of anime titles, and rate shows based on a scale from 1-10[1]. The motivation behind choosing this topic is due to a shared common interest in anime as well as prior familiarity with MyAnimeList site functionality. The knowledge and results gained from the analysis of this data set may be beneficial for future personal decisions regarding choosing an anime to watch.

Based on certain features from the data set – such as anime titles, genres, and ratings, to name a few – we aim to discover and answer interesting questions such as:

- Which genres (of anime) contribute to ambiguous users' watching decisions?
- How much do features such as rating, episode count, and anime type influence a user's viewing choices?
- Is the popularity of an anime or similarity to other users' profiles/viewing preferences more significant in impacting a user's preferences?

In addition to the questions listed above, we may also discover or explore other potentially interesting questions that may arise during the EDA portion or data mining/analysis phase of the project. Our project will aim to answer the questions above at minimum.

LITERATURE SURVEY

Prior work on this topic can be seen on popular anime streaming platforms such as Crunchyroll[2], which has an existing recommendation system in place which suggests various titles to a user based on the users' previous watch history on the site. The MyAnimeList site also provides a list of recommendation to users who have added anime entries to their profile, and these recommendations differ from those of Crunchyroll's since they also incorporate other users' unique feedback to support the recommendation[3].

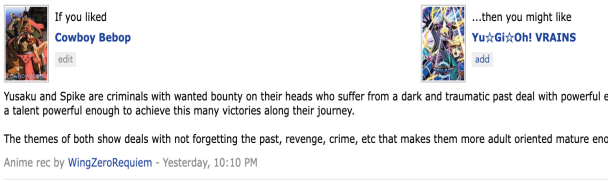

Figure 1: An example of recommendation with unique user commentary provided by MyAnimeList.net.

DATA SET

The data set that our project will be primarily using is titled Anime Recommendations Database, and it was

[1] Summary taken from https://myanimelist.net/about.php
[2] https://www.crunchyroll.com
[3] https://myanimelist.net/recommendations.php?s=recentrecs&t=anime

obtained from Kaggle.com. The data set can be accessed via this URL: https://www.kaggle.com/CooperUnion/anime-recommendations-database. The data set is comprised of two separate CSV files, Anime.csv and Rating.csv, and it contains information on user preference data from 73,516 users on 12,294 different anime titles[4]. The Anime.csv file contains various features (columns) describing anime titles, such as anime_id (the unique identifier assigned to an anime title by MyAnimeList), name, genre, type, number of episodes, rating, and members (the amount of users on the site who have added a particular anime to their profile). The Rating.csv file contains the features user_id (a randomly generated user ID), anime_id (the unique identifier assigned to an anime that the particular user has rated), and rating (the rating out of 1-10 that the particular user has assigned).

PROPOSED WORK

The data set obtained from Kaggle.com is already quite clean. We plan on performing an initial Exploratory Data Analysis on the data before the data pre-processing phase by creating various data visualizations. In the Rating.csv file, the "rating" column (which specifies the rating out of 1-10 that the user has assigned to a particular anime) contains the numeric value -1 if the user has added the anime to their profile but has not designed a rating, whether intentionally or not. To account for these values, we plan on either dropping these values entirely or substituting them with another constant value during the data pre-processing phase. In addition, it is important to note that in the Anime.csv file, there are several titles which belong to the same anime franchise (for example, it contains an entry for the first season of the anime "Haikyuu!!" as well as an entry for a special movie episode of "Haikyuu!!"). To account for these occurrences, we also plan to combine the entries (based on having the same title) which belong to the same anime franchise in the data pre-processing step, since these entries typically consist of 1 episode (as opposed to the typical number of 12 to 24 for most anime series) and could possibly skew the distribution of the ratings for a particular anime(s).

Our project will also aim to develop and train a proper classification model for the data set. A feature that we plan to incorporate that will differ from the prior work done that was mentioned in the Literature Survey portion is to allow for a two-way recommendation system in which user input can also be provided in addition to the existing user data in the data set. The output will be a result of shows that pass a percentage threshold of matching for one of a few models that will test different types of similarity. This can include multiple ways of abstracting the data (studio, art style, type of media). We will also be checking multiple types of models either clustering, decision trees, random forest, k nearest, linear regression, naïve bayes, or k nearest.

TOOLS

The tools which we plan on using for this project include:

- Jupyter Notebook (for formatting and code)
- Pandas, numpy, sklearn, scikitlearn (for calculations, EDA, data processing, and data analysis)
- Matplotlib, seaborn (for visualizations)

EVALUATION METRICS

The evaluation metrics which we plan to use to evaluate our classification model currently include, but are not limited to, are: cross validation, precision,

[4] From the description of the dataset found at https://www.kaggle.com/CooperUnion/anime-recommendations-database

accuracy, F1 score, recall, and R-squared value. The metrics will be used to help us decide which model is effective and drive the best recommendation. The problem that we have really identified with inferior recommendations is due to not having a high enough recall and precision. Recall tells us how often we are correctly identifying a good recommendation while precision will tell us how the model performs over the whole data set. The f1 score will give us a representation of recall and precision together. Ideally, we want to maximize all the scores with focus on the precision. The focus on precision allows us to make the claim of a "better" recommendation system.

MILESTONES

At the time of this proposal, we plan to have at least completed the Exploratory Data Analysis and data-preprocessing steps by the end of Week 10 (March 18). In addition, we hope to be able to finish a majority of the coding portion and model training by the end of Week 11 (April 1).

REFERENCES

[1] https://myanimelist.net/about.php
[2] https://www.crunchyroll.com
[3]https://myanimelist.net/recommendations.php?s=recentrecs&t=anime
[4] https://www.kaggle.com/CooperUnion/anime-recommendations-database