Anime List - Dataset Preparation

Andreia Rodrigues
up201404691@fe.up.pt

Francisco Queirós
up201404326@fe.up.pt

Miriam Gonçalves
up201403441@fe.up.pt

19 de Outubro de 2018

Abstract

Anime refers to "animation" and it represents all animation produced by Japan. It has a distinct look-and-feel compared to western animations and, over the last forty years, it has become an international phenomenon, attracting millions of fans and being translated into many languages.

With over 100 different anime tv shows and movies being released every year, anime fans started to gather information about their favorite shows and creating large databases to share with the rest of the community.

1 Introduction

Anime is an abbreviation of the word "animation" and it is used by the Western culture to describe a Japanese-style animated film or tv show. It is characterised by the characters having huge eyes, bright colored hair and exaggerated emotional expressions and gestures.

Japan began producing animation in 1917, but *anime* only started to become famous after the 60s due to the creation of television, that had a crucial role on making these japanese characteristic films an increasing trend.

They usually originate from *manga* (comic books), visual novels, light novels or video game adaptations.

These animations encompass a diverse range of storytelling styles, from horror to romance and everything in between, being a popular form of entertainment for adults as well as for younger audiences. *Anime* often covers more serious topics other than the typical cartoon.

Nowadays, with the growth of *anime* popularity and the number of animated films being produced every year, *anime* fans gather in online platforms where information about *anime* and its reviews are collected and can be accessed, allowing users to interact with each other, share what animations they are interested in and keep track of what they have watched, are watching and want to watch in the future.

These online platforms gather their own data separately and there isn't a centralized place where users can have an overview of an anime and the rating it is given in each of them. To counter this we intend to get and process all the necessary data from the most popular *anime* rating platforms, find a suitable way to store it and develop intuitive methods for users to search for *anime* they are/might be interested in.

2 Background

Over the years, with the major increase of *anime*'s popularity, several platforms have been created for sharing information about these animated movies and tv shows. The need for specific websites comes from *anime* having a very distinct style compared to regular animations that doesn't really fit the reality of popular movie and show rating websites like *IMDb* or *Rotten Tomatoes*.

The most popular and complete websites are "*MyAnimeList*", "*Anime News Network*", "*AniList*" and "*Kitsu*". They all complement each other in terms of information available and ratings attributed. Each of these websites allow users to create accounts and keep track of what *anime* they have watched, are watching, want to watch and discuss and review each *anime*. *IMDb* has recently started to include the most popular and better rated *anime* in their lists.

3 Information Retrieval

The dataset used as our basis comes from *Kaggle*, a platform for predictive analytics competitions in which companies and researchers publish datasets while statisticians and data miners compete to produce the best models for predicting and describing the data. *Kaggle* holds many datasets that may vary on the subject and the quantity of data.

For the dataset we are using, we only retrieved the file containing the general information about *anime*. It has the name "*anime_cleaned*" and contains 6.03 MB of data in a comma-separated values (*.csv*) file. The information in this dataset was collected from the most popular *anime* website "*MyAnimeList*", where *anime* fans share information about each *anime*, discuss about it and rate the ones they have watched while also keeping track of what they are watching, have watched and want to watch.

The second dataset used comes from *IMDb*, the largest movie, cinema and tv shows database. It is one of the most relevant platforms in the cinema industry with over 4.7 million titles (movies and tv shows), 8.3 million personalities (actors, producers, etc) and 83 million users registered. Users can give a rating score from 1 to 10 to any existing title in the database. The website provides vast and complete information about each title as well as the average rating score.

The data files we are using are called "*title.akas*", a 171 MB file with the most basic information about the titles,"*title.rating*", a 14.1 MB file that connects the *ids* of titles provided in the first file with the respective average rating and "*title.basics*", a 431MB file that connects the titles *ids* to some other additional information like the type/format (movie, short, animation, etc). All these files are tab-separated files (*.tsv*).

Lastly we will use the data provided by the *Kitsu API*, an *API* that retrieves data from the *Kitsu* website, one of the most popular anime websites, similar to "*MyAnimeList*", where the community rates the shows they have watched. This *API* is well documented and the data is returned in *JSON* format and can be filtered beforehand on the *API* call.

4 Data Preparation

The first step on the data preparation task was to clean and organize the basis dataset, as it had some repeated information and some that we didn't require, to get a better understanding of the data structure.

To do that, we used *OpenRefine*, a tool for working with unstructured large data sets, allowing to clean up data, transform it to other formats and extend it with web services and external data. It allows the user to filter by each column, changing which rows should be displayed, remove columns, amongst other useful things.

We used this tool to remove the columns with the redundant data and the ones that were irrelevant for the purpose of this project. We also transformed the *.csv* file into a table like representation, so that we could see more clearly what data was correlated and what data was missing. From here we identified the complementary information that we could use for this dataset and the other datasets/*API*s we could use to get it.

To enrich our data we started by importing the data provenient from the "*MyAnimeList*" *Kaggle* dataset and the *IMDb* title dataset onto a *SQL Server* database and filtering the data that matched on the "title" attribute, since there was nothing else that related both information. If more that one result for a particular *anime*'s title show up, we'll filter those titles for those that have animation as a genre. *Anime* doesn't exist as a genre in *IMDb*, but *anime* implies being an animation.

To relate the rating data from *IMDb* to the respective *anime*, the ids of each dataset referencing that title are used.

There was still some information that we found useful missing (the synopsis and trailer link for example) so we decided to gather the information from another popular *anime* community website called *Kitsu*. The relation between the *anime* data we already had and the

Kitsu data would also be done by the title, since there was no other way to connect both.

5 Data Limitations

There were some limitations relative to the crossing of the data between the different datasets since the only way to relate data was using the *anime* title and that can cause some restrictions if the field isn't coherent.

Regarding the data being updated, the only issues that can be faced are regarding the *MyAnimeList* data. Since the website's *API* was taken down, the dataset is only updated by the user that submitted this data to *Kaggle* and we don't know how frequent that will be done. Therefore, it can have some information that isn't fully coherent with what the website is showing at the moment.

The *IMDb* dataset is updated by their official website so we will assume that no problems will show up and that the updates happen regularly.

The *Kitsu* data will always be updated because it will be coming directly from their *API*.

6 Data Model

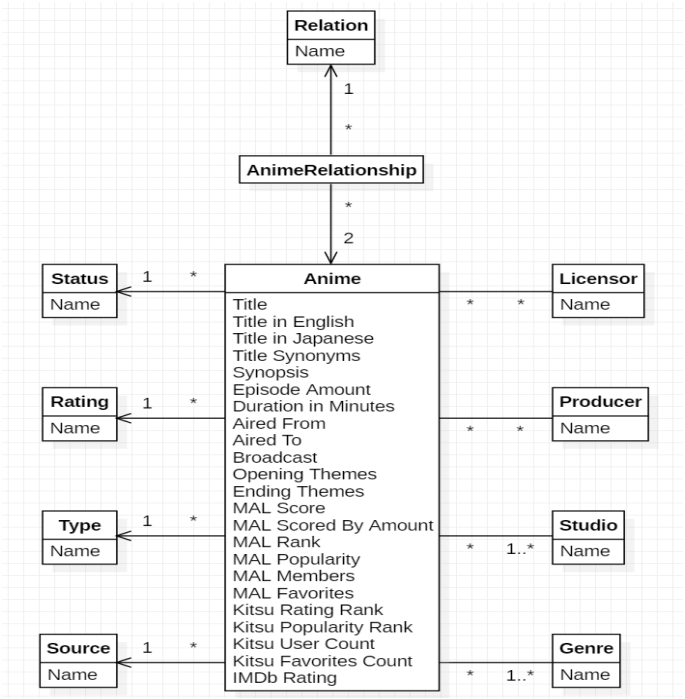

***Fig. 1** - Anime List Data Model.*

- **Title/ Title in English/ Title in Japanese/ Title Synonyms** - These are names given to the *anime*. The names are self-explanatory.
- **Synopsis** - This contains a synopsis of the *anime*.
- **Episode Amount** - Amount of episodes in the *anime*.
- **Duration in Minutes** - How long a typical episode of the *anime* is, in minutes.
- **Aired From/To** - Dates when the *anime* started to be aired and, if it has already ended, when it stopped being aired.
- **Broadcast** - Day of the week and time of day the *anime* is aired.
- **Opening/Ending Themes** - These represent the name of songs that are played at the start and end of episodes of the *anime*.
- **MAL Score/ Scored By Amount/ Rank/ Popularity/ Members/ Favorites** - These values represent different aspects of how users from *MyAnimeList* share their fondness for this *anime*. The "score" is the average score that the amount of users on the "scored by amount" field rated the *anime*. The "rank" is based on the top rated *anime* page. The "popularity" is based on how many people have added the *anime* to their lists (the "members" field). The "favorites" field is related to how many people have included this *anime* in their favorites.
- **Kitsu Rating Rank/ Popularity Rank/ User Count/ Favorites Count** - These values represent different aspects of how users that were sourced by *Kitsu API* share their fondness for this *anime*. The "rating rank" represents the place of the *anime* in top ranking (how well it's scored). The "popularity rank" represents the ranking of the popularity of the *anime* (how many users know/watch/like the *anime*). The "user count" represents how many users interact with the *anime*. The "favorites count" represents how many users put the *anime* as a favorite.

- **IMDb Rating** - This represents the score that *IMDb* has for this *anime*.
- **Status** - Represent the status of the *anime*, for example, if it's still being aired, has finished or has been cancelled, among others.
- **Rating** - Content rating that measures its suitability for different audiences. One example is PG-13.
- **Type** - The format of the *anime*. Either for TV broadcasting, a movie, among other formats.
- **Source** - Type of content from which the *anime* was based.
- **Licensor** - Group that acquires permission to transmit the *anime* in other regions.
- **Producer** - Group that oversees the product during development and gives guidance.
- **Studio** - Group that participates in directly making the *anime* itself.
- **Genre** - Genre of the *anime*, for example, Comedy.
- **Relation** - Some *anime* relate to another *anime* in some way. For example, sequel, prequel, or telling a side-story, among other relations.

7 Domain Model

The domain model shouldn't be at all different from the data model. We don't have other entities not mentioned in the data that would be relevant to understand or draw more complete conclusion from the data.

8 Data Visualization

With this data we can make queries that go from simple:

- "What is the *anime* with the id 123?"
- "What are the *anime* with an attribute with a particular value? (e.g. Rating as PG-13)"
- "What are the studios/producers/etc. for this *anime*?"

To more complex queries:

- "How many comedy *anime* has a certain studio made?"
- "What is the most popular *anime* of a certain producer?"
- "What is the average score of the *anime* of a particular licensor?"

These queries can then later be used with full-text search. This will allow to query for keywords in big text fields like the synopsis. Another interesting idea would be to allow simple text as the query but match several fields. For example, if the user queries "PG-13", *anime* with "PG-13" rating could be returned to the user.

9 Conclusion

For this milestone it was required to look for appropriate datasets and determine what tools would be necessary to process and refine the data collected.

After analysing the existing datasets about *animes* it was important to understand how to cross data together in order to enrich the initial dataset chosen. This step was the hardest, since the cross between data was made by titles, that can be incoherent between datasets, and not by a predefined *id*. The final result is a large dataset with complete and coherent data about *animes*, which will be queried to get certain subsets info and related informations between others.

References

[1] *Kaggle: Your Home for Data Science*, https://www.kaggle.com/, 2018. [Online, accessed 25-September-2018]

[2] *Kaggle* https://en.wikipedia.org/wiki/Kaggle, 2018. [Online, accessed 26-September-2018]

[3] *My Anime List | Kaggle*, https://www.kaggle.com/azathoth42/myanimelist, 2018. [Online, accessed 26-September-2018]

[4] *My Anime List*, https://myanimelist.net/, 2018. [Online, accessed 26-September-2018]

[5] *IMDb Datasets*, https://www.imdb.com/interfaces/, 2018. [Online, accessed 28-September-2018]

[6] *Openrefine : A free, open source, power tool for working with messy data*, http://openrefine.org/, 2018. [Online, accessed 10-October-2018]

[7] *Kitsu API*, https://kitsu.docs.apiary.io/, 2018. [Online, accessed 12-October-2018]

[8] *Wikipedia* page *Anime*, https://en.wikipedia.org/wiki/Anime, 2018. [Online, accessed 13-October-2018]

[9] *Wikipedia* page *History of Anime*, https://en.wikipedia.org/wiki/History_of_anime, 2018. [Online, accessed 13-October-2018]

[10] *What is Anime? A Brief History of Anime Genres, Culture and Evolution - The Daily Dot*, https://www.dailydot.com/parsec/what-is-anime/, 2018. [Online, accessed 14-October-2018].

[11] *A Brief History of Anime - Thought Co*, https://www.thoughtco.com/brief-history-of-anime-14497, 2018. [Online, accessed 14-October-2018]