Frequent Itemset Mining for MyAnimeList data

Keven Quach (9250982)

30 April 2021

1 Introduction

This report shows descriptive visualization and the application of frequent itemset mining for MyAnimeList data in order to receive an exemption from the course *Business Intelligence* at Utrecht University. It is shown that experience in the subject was gained through this additional assignment. The code can be found under https://github.com/beld78/MyAnimeList-Analysis.

MyAnimeList is a website where users can manage their personal list of anime they plan to watch, are currently watching, dropped or completed. The dataset[1] was collected in 2018 and consists of three files: users, anime and anime users have added to their list. The goal of the assignment is to retrieve meaningful rule associations between anime to make recommendations for further anime to watch. Therefore, the file of primary concern is the last one mentioned where each row represents an entry of a user and an anime, which will be referred to as animelist data. However, some descriptive analytics can also be found on the other datasets.

2 Methods

2.1 Preprocessing

Preprocessing of the animelist data is necessary in order to apply frequent itemset mining. Columns that are not relevant for the task were excluded. Additionally, only entries that have the status *completed* were considered. Other categories include *watching, on hold, dropped, plan to watch.* Since the goal is to get recommendations for an anime it doesn't make sense to include these entries.

Afterwards, the anime id was replaced by the actual title and the data was sorted by the last timestamp update from earliest to latest. Then, the rows were merged on the username in the order of the timestamps to properly recreate the actual viewing order as can be seen in Table 1 and Table 2. Note that Table 2 only shows a single row and line breaks were added for readability. Only rows with more than 10 items were included since these users were not actively using their list and can be seen as noise.

[1] https://www.kaggle.com/azathoth42/myanimelist

username	anime_id	my_last_updated
example_user	1	1159392517
example_user	30240	1240531353
example_user	20	1346574421

Table 1: Before preprocessing (3 rows)

username	anime_id
example_user	Cowboy Bebop|| Prison School|| Naruto

Table 2: After preprocessing (1 row)

For the anime data, only ones that were *finished airing* were considered, as the variables *rating, score, scored_by, rank, popularity and favorites* tend to be low until the anime is finished.

For the users, only ones that had more than 10 completed anime were considered in order to be aligned with the animelist data.

2.2 Further filtering

Further filtering of the association rules that were mined from the animelist data was necessary since it is quite common that anime have sequels or corresponding movies. While these obviously have a high confidence and support, they are not interesting in the sense of a recommendation. Therefore, these were filtered by checking whether a word on the left-hand-side of the rule also appears on the right-hand-side. Common fillwords and symbols were excluded. For example, the rule *Naruto → Naruto: Shippuden* would be removed. Some sequels do not fall into this pattern and are therefore not removed. An example would be *Bakemonogatari → Nisemonogatari.*

Only rules of length 2 were considered for two reasons. First, if the further filtering approach is used, most rules that are longer than 2 will be removed since they most likely contain a sequel or movie on either side. Second, since the goal is to create a recommendation, rules of length 2 are the most interesting ones.

2.3 Two use cases for association rules

The first use case is to extract the top N rules of all possible combinations and gain further insights into the overall animelist data.

The second use case is to mine the rules for one specific title. In that case it is interesting to have all rules where the title is either on the left-hand-side or the right-hand-side. Therefore, a function was created that takes a title, animelist data as a transactions object, support and confidence. The number of retrieved rules highly depends on the parameters support and confidence and is variable depending on the title.

3 Results

Descriptive visualization of both anime and user data was applied to get a better understanding of the dataset. The following two subsections show the respective results.

3.1 Anime

Figure 1 shows the distribution of total number of scored anime grouped by age ratings. It becomes clear that the vast majority of anime watched has a minimum age rating of PG-13. This is to be expected as the most popular anime like Naruto or One Piece are shonen[2] which are usually rated PG-13.

From Figure 2 the numeric variables are shown in a correlation matrix. The rank has a high correlation with popularity and score. Since a better rank is achieved by a lower number, the correlation with the score is negative. Episodes has a slight correlation with score which indicates that long-running series like One Piece tend to have a slightly higher score than the average anime[3].

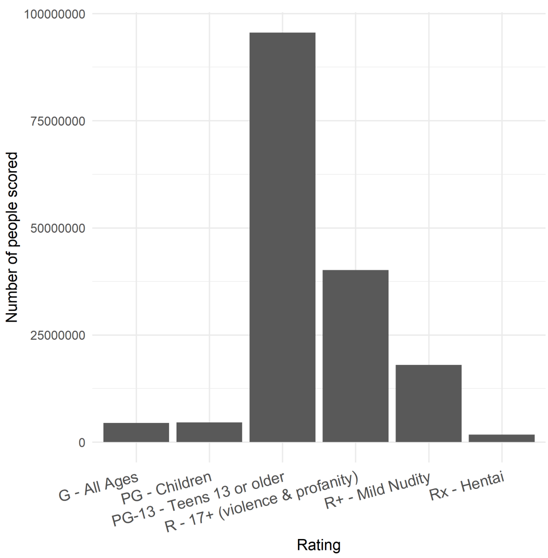

Figure 1: Number of people scored by age rating

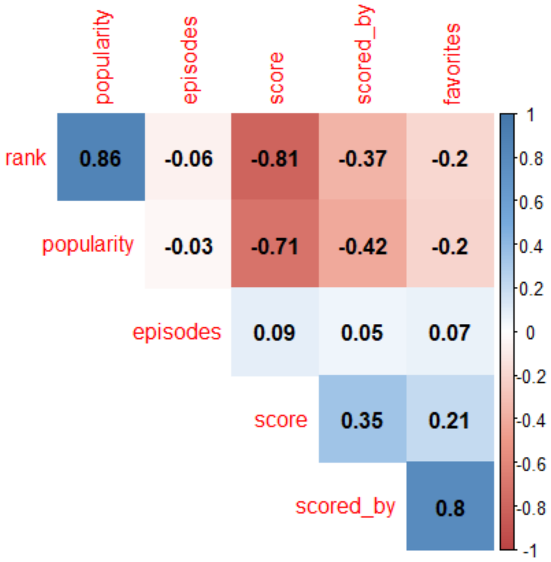

Figure 2: Correlation matrix of variables

3.2 Users

Figure 3 shows the distribution of the average score grouped by gender. There are no extreme deviations within the different groups. The overall average score is 7.68 and most of the scores are within the range of 6 to 10.

Figure 4 shows the cumulative distribution function of the number of completed anime. Nearly all of the users have less than 1000 anime in their list. However,

[2] Japanese comics marketed towards young teen males between the ages of 12 and 18

[3] The normal number of episodes for an anime is 12 (one season) or 24 (two seasons)

there are 1814 users with more than 1000 anime completed. These observations are not considered noise as there could very well be people that have watched this amount[4].

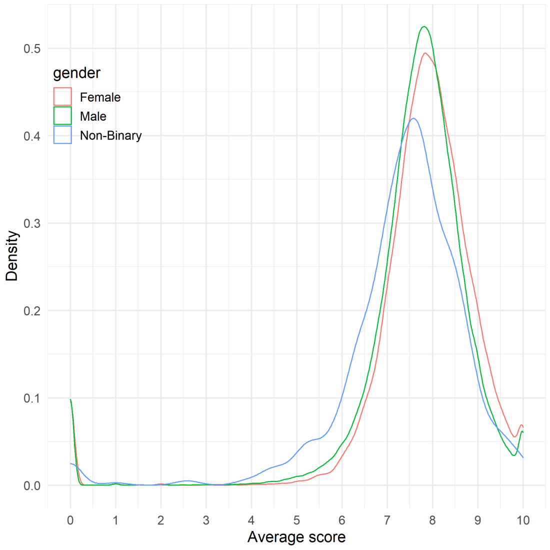

Figure 3: Distribution of average scores

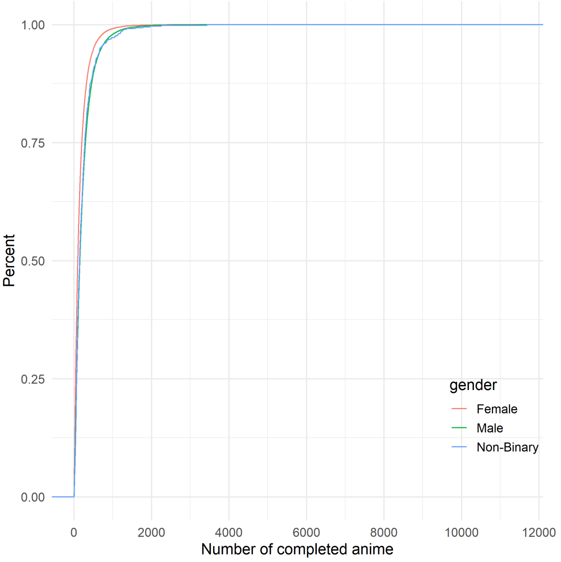

Figure 4: CDF of completed anime

3.3 Animelist

Figure 5 shows the top 20 anime completed in the dataset. This shows that there are anime that are frequently watched by many people. *Death Note* was completed by more than 66.43% of all users. Afterwards, there is a large drop to 52.79% for the second most frequently completed anime. However, frequency does not take the scoring into account. For example, *Another* is currently ranked at 1487 with popularity at 46[5]

Figure 6 and Figure 7 show the rules for all transactions with the following parameters: support = 0.1; confidence = 0.8; maximum and minimum rule length = 2. Plotted are their support, confidence and lift before and after filtering the rules. Most of the rules with confidence close to 1 are filtered out and 213 rules in total were removed.

Table 3 shows the top 10 rules ordered by lift. Number 1 to 3 are sequels as previously mentioned. The following 7 rules are correct recommendations and have the same genre on both sides. Rule 8, *Koe no Katachi* and *Kimi no Na wa*, are both regarded as some of the best anime movies and are often recommended together.

[4] Specials, extra episodes, movies and similar content all count as unique entries

[5] Retrieved at 27.04.2021: https://myanimelist.net/anime/11111/Another

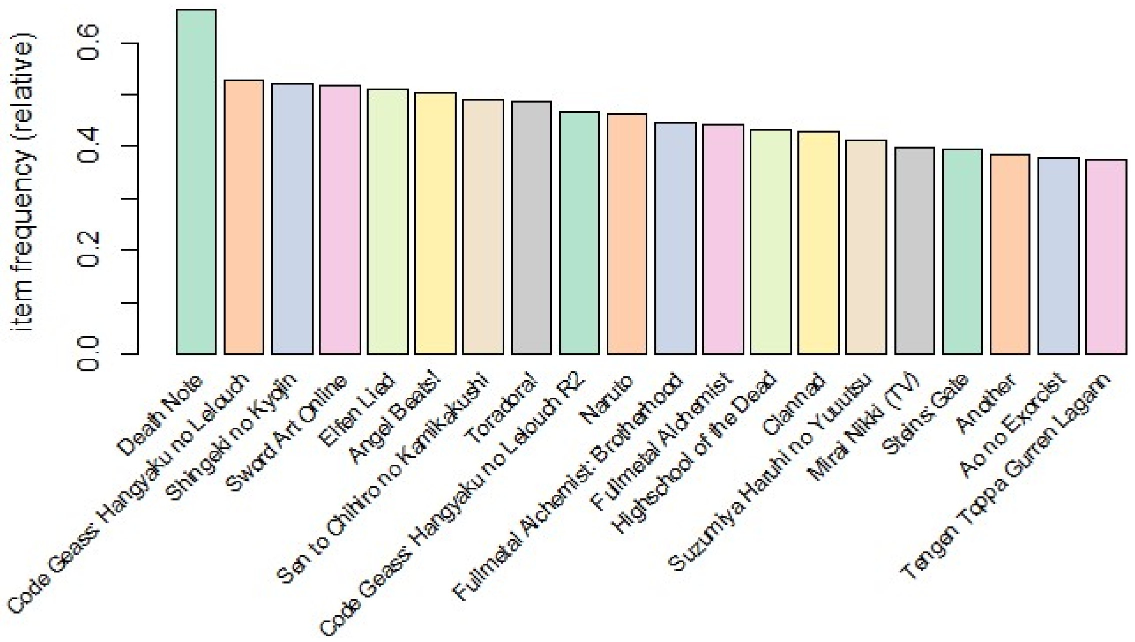

Figure 5: Most frequent anime

#	Left-hand-side	Right-hand-side	Lift
1	{Hanamonogatari}	{Nisemonogatari}	4.28
2	{Monogatari Series: Second Season}	{Nisemonogatari}	4.21
3	{Nekomonogatari: Kuro}	{Nisemonogatari}	4.11
4	{Rakudai Kishi no Cavalry}	{Dungeon ni Deai wo Motomeru no wa Machigatteiru Darou ka}	3.51
5	{Kono Subarashii Sekai ni Shukufuku wo! 2}	{Re:Zero kara Hajimeru Isekai Seikatsu}	3.28
6	{Saenai Heroine no Sodatekata}	{Yahari Ore no Seishun Love Comedy wa Machigatteiru.}	3.18
7	{Dakara Boku wa, H ga Dekinai.}	{High School DxD}	3.14
8	{Koe no Katachi}	{Kimi no Na wa.}	3.11
9	{Hagure Yuusha no Aesthetica}	{High School DxD}	2.97
10	{Mob Psycho 100}	{Boku no Hero Academia}	2.96

Table 3: Top 10 rules ordered by lift

Table 4 shows all rules for *Prison School* with parameters: support = 0.01; confidence = 0.7. Rule 2-11 have quite popular anime on the right-hand-side. Rule 1, however, is a good recommendation since both anime are not too popular and have the same target audience. The lift is also much higher than the other rules, indicating a high positive effect.

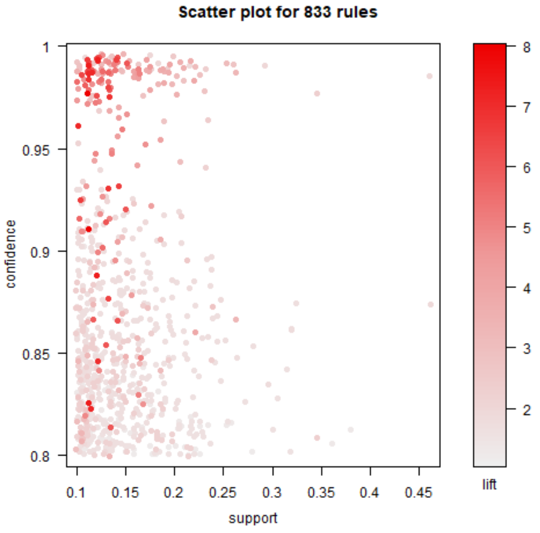

Figure 6: Unfiltered rules

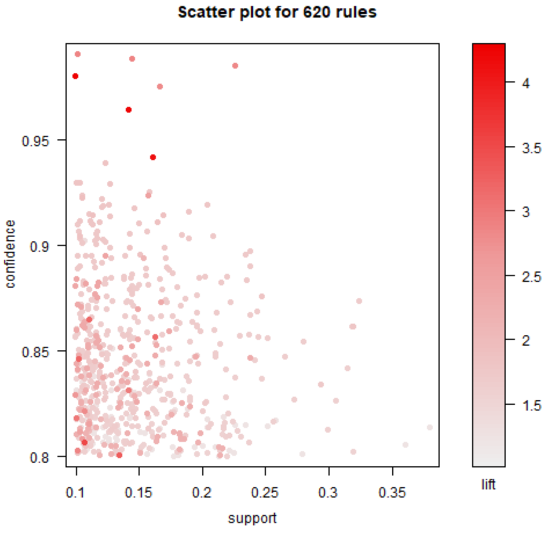

Figure 7: Filtered rules

#	Left-hand-side	Right-hand-side	Lift
1	{Sin: Nanatsu no Taizai}	{Prison School}	4.35
2	{Prison School}	{One Punch Man}	2.23
3	{Prison School}	{Tokyo Ghoul}	2.16
4	{Prison School}	{Noragami}	2.15
5	{Prison School}	{No Game No Life}	2.13
6	{Prison School}	{Mirai Nikki (TV)}	1.88
7	{Prison School}	{Highschool of the Dead}	1.72
8	{Prison School}	{Shingeki no Kyojin}	1.70
9	{Prison School}	{Sword Art Online}	1.63
10	{Prison School}	{Angel Beats!}	1.50
11	{Prison School}	{Death Note}	1.21

Table 4: All rules ordered by lift

4 Evaluation

There were quite a few hurdles in the assignment. As usual in data science, the data preprocessing was a large part and proved to be difficult. While it would have been easier to use a dataset that is already in a transaction format and perfecty suited, it helps to use such a dataset to gain a deeper understanding of different implications. For example, the filtering of sequels and related movies is something that is unique to this kind of data. The filtering could also be further improved to check for different word stems. Series like Bakemonogatari could be then also excluded.

It is also questionable how useful frequent itemset mining as a recommendation system is. The rules have a strong bias towards anime with a higher frequency as can be seen in Table 4. Most importantly, frequent itemset mining only takes frequency into account. Especially for a recommendation, it would be preferable to include user scores.