Analysing Steam Game Library and User Behaviour

Ugur Ali Kaplan
Matrikelnummer 6031686
ugur.kaplan@student.uni-tuebingen.de

February 7, 2022

Abstract

We have analyzed various video games in the Steam Library, using the Steam Video Games Dataset. We used the purchase information to visualize similar games with t-SNE. We compared Jaccard index and Euclidean Distance metrics in terms of how they affect the outcome of the visualization. We observed if the visualizations of the t-SNE were faithful to the Jaccard Index or not. Our code and the higher resolution version of the figures used in the report are published on GitHub.

1 Introduction

In this report, we will be reviewing the user and video game interaction on the popular video game platform Steam [1]

In this project, we have done the following:

1. We have measured the similarity of games, using a popular pairwise function called Jaccard Index [4]. We observed if the similar games, according to the Jaccard Index, reflected reality and our expectations.
2. We have visualized our findings, using t-SNE [6].
3. In our visualizations, we compared using different distance metrics for t-SNE and checked if that made any difference.
4. We compared if the t-SNE visualizations grouped similar games, according to the metric it uses.

In addition to these, we have tried using the playtime instead of the purchase information, and using Funk SVD [3] for detecting similarities with higher performance and failed. Details of that failed experiment can be found on GitHub.

2 Methods

2.1 Game Similarity with Purchase Data

There are various methods we can use to measure similarity. The Jaccard index is a popular similarity metric. It is used to calculate the similarity of sets. Let S_1 and S_2 be different sets defined over the same field.

$$Sim(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

To use this for the players and games, we treat the games as sets, and the elements of the set are the players that own the game. If a game XYZ is owned by Alice, Bob, Charlie, we represent the game as:

$$S_{XYZ} = \{Alice, Bob, Charlie\}$$

We could also want to measure the similarities between users, and from a corporate point of view, that could be profitable but there is also the privacy aspect. When we use user-user similarity, there is no way to explain the reason for a recommendation, as it would require discussing one user's interests with an unrelated party.

2.2 Visualization of Games

To visualize the games, we have used t-SNE as it preserves the local structure and most similar games will be clustered together. In [5], it is suggested that using PCA initialization preserves the global structure better. Thus, we used PCA initialization. We have experimented with different values of perplexity, and used one of the options that had more distinctive clusters. Finally, we use both Jaccard and Euclidean distance as metrics and observe the differences in the clusters in the generated plots.

3 Data

We are using the Steam dataset released by Tamber [2]. It has 200.000 entries with 12.393 unique users and 5.155 unique games. 70.489 of these entries are of playtime data, and the rest 129.511 entries are the purchase data.

4 Results

4.1 Game Similarity

For this part, we have dropped users that own less than 20 games, as suggested by GroupLens [4]. In the end, we got a square similarity matrix with 1.197 rows and columns, each corresponding to a game.

To check if the algorithm worked, we used manual inspection of some games we are familiar with, which can be seen in table 1. The first row is the most similar game and is there for sanity check as all the players should be the same. As we go down, we get further away from the game we are interested in.

Mass Effect	**Skyrim**	**XCOM Enemy Unknown**
Mass Effect	Skyrim	XCOM Enemy Unknown
Mass Effect 2	Skyrim - Dawnguard	XCOM Enemy Within
Bioshock	Skyrim - Hearthfire	FTL Faster Than Light
Borderlands	Skyrim - Dragonborn	BioShock
Far Cry 2	Skyrim High Resolution Texture Pack	BioShock Infinite

Table 1: Similar games according to Jaccard Index

Our findings are in line with our expectations, games from the same franchise are together. For Mass Effect, other fast-paced action-themed games turned out to be similar. For Skyrim, Skyrim mods and DLCs are similar. For XCOM, sci-fi and space-themed games are similar.

4.2 Visualization

We used Euclidean distance and Jaccard Index on the set representation. Visualization results can be seen in fig. 1. When we use Euclidean distance, we get distinct clusters. When we observe these clusters, we see that they are meaningful. The ball-shaped big cluster in the top left consists of multiplayer games. Other populated but rather small clusters are composed of games from the same franchise, such as Mass Effect and Mass Effect 2. So, even though we used Euclidean distance as our metric, it got similar results to what we get from the Jaccard distance.

However, when we use Jaccard for visualization purposes, we saw that it is harder to detect clusters, and similar games according to the index are scattered across the plot.

We see that with Euclidean distance, points are scattered between -50 and 60 in the x-axis, and -80 and 80 in the y-axis. However, with the Jaccard index, the points are scattered between -3.5 to 2 in the x-axis, and -3 to 2.5 in the y-axis. So, we think this is caused by the famous "curse-of-dimensionality". When we use the Jaccard index, every point is somewhat close in the space but using Euclidean distance in 1.197 dimensions helped the visualization process.

5 Conclusion

We have seen Jaccard index can be successfully used to detect similarities between sets, or things we can represent as sets. t-SNE is heavily affected by different

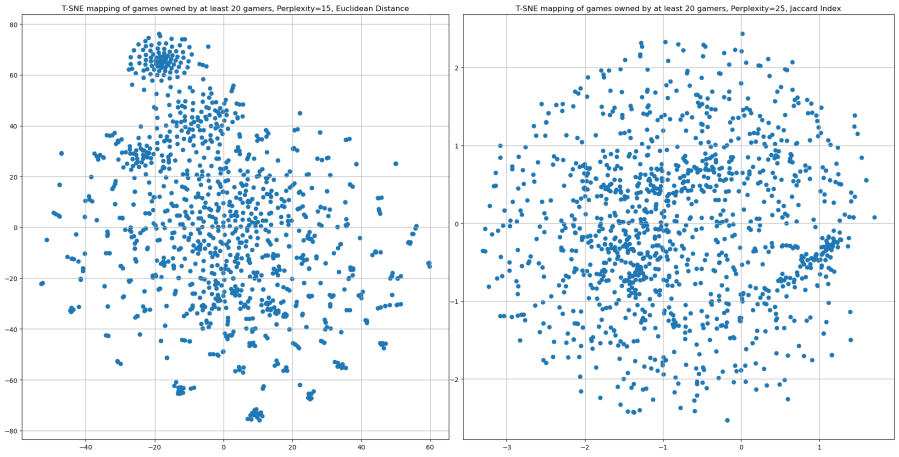

Figure 1: t-SNE visualization with Euclidean Distance (Left) and Jaccard Index (Right) as distance metrics.

choices of metrics. When we use Jaccard Index for the metric, t-SNE didn't necessarily preserve the local structure. So, even though the Jaccard index works well to find similar games, for visualization purposes using Euclidean distance proved to be more rewarding.

References

[1] Steam Community, Feb 2022. URL https://steamcommunity.com. [Online; accessed 6. Feb. 2022].

[2] Steam Video Games, Feb 2022. URL https://www.kaggle.com/tamber/steam-video-games. [Online; accessed 6. Feb. 2022].

[3] Netflix Update: Try This at Home, Feb 2022. URL https://sifter.org/simon/journal/20061211.html. [Online; accessed 6. Feb. 2022].

[4] Kim Falk. *Practical Recommender Systems*. Manning Publications. ISBN 978-1-61729270-5. URL https://learning.oreilly.com/library/view/practical-recommender-systems/9781617292705.

[5] Dmitry Kobak and George C. Linderman. Initialization is critical for preserving global data structure in both t-sne and umap. *Nature Biotechnology*, 39(2):156–157, Feb 2021. ISSN 1546-1696. doi: 10.1038/s41587-020-00809-z.

[6] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.