Hybrid Boardgame Recommender System
Sayan Patra
May, 2018

1 Introduction

Gaming is and has always been an important part of my life. Occasionally maybe more than being a successful statistician. Even being absolute garbage at Dota2, after playing it for almost 5 years (don't ask my MMR), does not curb my enthusiasm to queue up for a ranked match at the end of a long day. And what can be a better weekend party than a marathon 4 hour run of Terraforming Mars? This project, therefore in all sincerity, is an honest attempt to merge the best of both my worlds and hitherto it has been a terrific ride.
There is no denying that there has been a boom in the gaming industry for the past decade. As a result the quantity, quality, and diversity of the games are high. When there are literally hundreds of good options to choose from it becomes tricky to pick your next epic journey, a Boardgame recommender system can become handy. Because it is about the journey and not the destination and the fact that the quality of the journey is very much subjected to personal bias it is important for the recommender to be properly tuned to the user. As an example, where some of my "lesser-minded" friends might feel head over heels for Codenames I would rather spend that valuable time to save this world (and those friends) from the cruel grasp of Cthulhu in Arkham Horror. Moreover, it would be great if a recommender can go beyond the trending games and suggest games that are not popular recently if the user wishes. For those unfortunate souls like me who did not have the good fortune to enjoy Dungeons & Dragons in their youth, this would be a game changer, or meta breaker as one might say. Therefore for my Data Incubator project, I am proposing to build a boardgame recommender system which can break the meta and provides the user with a game category analysis along with the suggesting games.
There does exist a few online boardgame recommender systems, particularly the one by Quantic Foundry is quite impressive. Most of these use a collaborative filtering on the individual user-game rating to suggest similar games. While the massive library of user inputs cancels the user bias, it can only suggest games that are recent or very popular. Also, none of these recommenders output the underlying feature of your game choices. In this project, I build a hybrid recommender system using content-based filtering and metadata based similarity scores that can suggest both newer and older games based on the user's input and outputs a profile of the users choice of games. The user can use this profiling to look for similar games in this category or go out of his comfort zone and explore other genres. I also produce a clustering map of all the games in the BGG database which has at least 30 user ratings. The clustering shows the existence of 3 broad categories each pertaining to a very different type of board game experience. These can be roughly described as children/ party games, quick card games, and heavy strategy games. This broad clustering can be interpreted as an indicator of the subgroups of the population with vastly different choice in boardgames. By focusing on the features of the subclusters artists and designers can have a better understanding of what mechanics or aesthetics work for their target subpopulation. The feature targeting will enhance the gameplay in general, making the gaming enthusiasts happy, fetch more revenue for the artists, designers, and publishers, and generate more data for my algorithm. Therefore my algorithm works for both the consumer and the producer level.

2 Data cleaning

I used the dataset provided by Gabriele Baldassarre in Kaggle. The dataset is scrapped from the Board Game Geek (BGG) website and contains attributes and ratings for around 94000 games. The resulting .sqlite file was imported into R before being worked with. This particular dataset is quite famous among the community and there are several blogs and Kaggle pages that do exploratory data analysis to some extent. Though these analyses are unrelated and quite basic in nature, they cover a vast majority of topics and established the cornerstone of several of my decisions regarding data cleaning and modeling.
One of the main problems I faced was that a vast majority of games have very low user rating. This consists very old regional games, European games, and newer non-USA based games which did not receive much mainstream attention. This makes sense as the number of games published each year is growing exponentially from the later part of 1980 and therefore comprehensive database of all the games and their attributes have only recently emerged. However, it is very problematic to consider these games as user rating, though a very useful attribute, suffers from the personal bias of the

user for the game. If considering it can potentially bump up the relevancy of a game, especially when the number of the rating is low. Therefore we have chosen only the games which have more than 30 user ratings to make it statistically relevant. Ambiguous datapoints like a game with publication year 500 or nonpositive maximum playtime were also removed. We also omitted all the columns which were mostly empty. This leaves us with a rich dataset of about 15000 games consisting of both newer stomper like Gloomhaven and older masterpieces like Catan which are at least known to some extent.

3 Model

Aside from handling a vast resource of data and making the best use of it the BGG dataset also presents us with some statistical challenges. Because the pulled dataset is aggregated over all the users, i.e. we can not get user-boardgame interaction data on an individual level, one can not do collaborative filtering. This forces one to understand the context and effect of the variables and build a supervised learning method for the recommender system. The excellent analyses provided by Gallen Ballew illustrates that even though many of the features have linear independence they fail to correctly capture the rating of a board game. This is mainly due to the nature of the variables itself. Variables like how many people want or wish a game do not say much, and variables like total owners, number of people interested or want to trade are either self-explanatory or requires a stochastic time series modeling. More recent games are also rated highly due to increased user population on BGG. Therefore we focus on a recommender system that instead of crunching numbers, looks at the textual description of the dataset.

4 Description based recommender

The personalized recommender engine we are building is based on the textual data of the games in the dataset. This content-based filtering system builds a TF-IDF matrix based on the description, artist and designer of the game. The idea is that this subset of the data contains just enough of the information about the genre, world, and mechanics of the game to create a soft clustering. As an example, a card-based political game like 7 Wonders contains the words card and politics but does not specify that it is a deck building game. Similarly, Catan mentions it involves both card and tiles but omits the word area control. The absence of statistics like average user rating and the number of users who rated the game ensures that this is free of any user bias. This enables the algorithm to suggest older and nonpopular games. There is an argument to be made about treating artist and designer data as categorical. However, many a time a person can be both designer and artist of a game and there this method favors the person. We did not include publisher data as it was very difficult to parse, mostly due to the difficult names of the northern European gaming companies.

The implementation of this method uses sparse matrix calculation package of R and is quite fast. The recommender system first aggregates the data for each game and preprocesses it to remove any digits and string of characters that are not sensible. It also removes spaces between names and any special characters. It then removes common stemming and stopwords in English and an additional custom-made list of stop words particular to gaming to create the dictionary. The tf-idf matrix is calculated from the bi-gram of the cleaned vocabulary. The algorithm subsets the matrix based on the game title(s) we provided, calculates a weighted average based on the weights provided by the user (if any) and returns the top 20 results. We found out that this version favors expansions and sin offs heavily. Even after we removed all the games marked as an expansion from the dataset, this remains true. As an example, a search title of Catan, Ra, and Tikal returned 9 results which were Catan spin-offs or different versions of the base game. Additionally, the system completely ignores the different gameplay mechanic of Tikal and Catan. It becomes evident we need to incorporate the mechanic and category of the game into the system.

5 Metadata based clustering

The BGG dataset contains two columns named category and mechanic which describes the gameplay aspect of the particular board game. Though the two variables might seem similar, they describe two very different features of a game. ”Catan” is a card game by category and area control by mechanism while ”Sentinels of Multiverse” is a Card game by category and deck building, cooperative by the mechanism. In our algorithm we build a sparse binary matrix of all

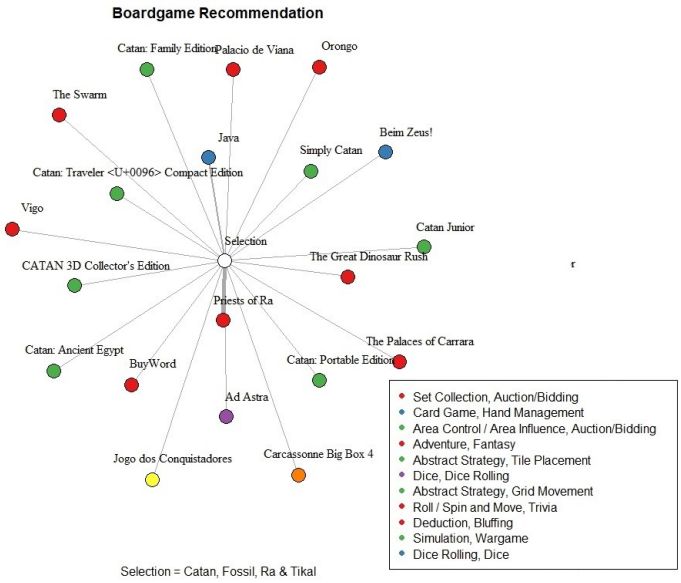
the categories and mechanisms and then apply the tsne algorithm, using Rtsne package, to create a low dimensional representation of the whole dataset. We then use a hierarchical clustering with wards method to obtain the clusters. The data was first sampled into several 1000 length chunks and we clustered these quickly using medoid clustering (PAM package in R) to figure out approximately the optimal number of clusters, which was set to 50 for the entire dataset. The reason we did not use medoid clustering for the entire dataset is that pam suffers heavily from the curse of dimensionality, and our dataset had over 150 features. The dimensionality reduction algorithm of the tsne makes sense theoretically and also performed fast to produce logical clustering in our case. We do not have a quantitative metric to judge our machine's performance so this will have to be done qualitatively. The resulting cluster descriptions are informative in their own rights, as it shows that even though card games and dice games are part of almost every board game these are not as dominant as before. The current meta is diverging towards either mechanically heavy games like Scythe or easy party games like Codenames. Wargames are still a genre of its own.
In another version, say enhanced clustering, of the clustering we additionally considered publication year of the game as a factor variable, and maximum and minimum player number, maximum and minimum playing time and minimum age as ordinal variables. Because we are using mixed data type the Gower metric makes the most sense to calculate distance. However, as the Gower metric is unsuitable for non-normalized data we apply the Boxcox transformation to the ordinal variables apriori. After the dimensionality reduction using tsne, we perform a hierarchical clustering as before. The resulting clusters all have the additional variables as important cluster identifier and display a very interesting phenomenon. When the additional variables are considered 3 very distinct clusters emerge from the chaos. These can be roughly be categorized as children's games or easy party games like "Sushi Go", "Pit"; quick moderate difficulty games like "Karmaka", "Dominion" and mechanically heavy strategy games with longer playing time like "Zombicide", "Arkham Horror: The Card Game". A more in-depth analysis can discover the underlying variables and find patterns related to the subclusters.
The clustering is the most time-consuming step in our algorithm. However this clustering does not depend on the user input, therefore one can compute the clustering and store the cluster assignment of each game along with the cluster information beforehand. By pulling in the pre-computed cluster information in data fetching stage the recommendation becomes almost instant.

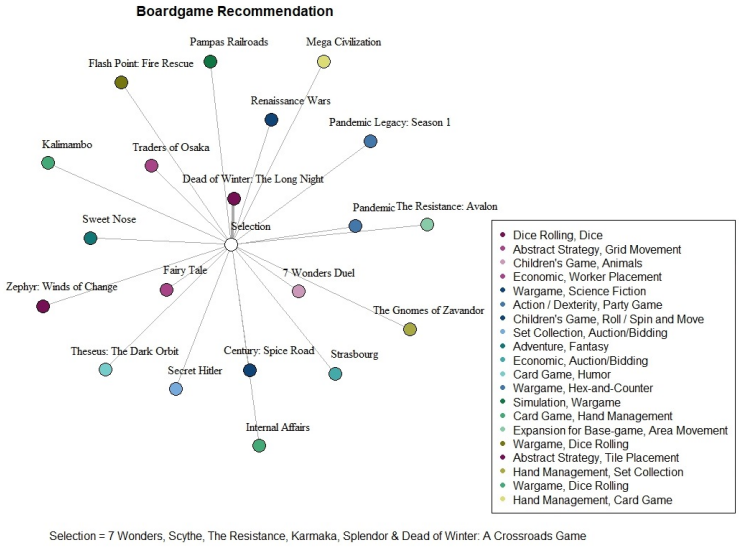
6 Hybrid recommender

Encouraged by the result of our clustering we decided to integrate our models. Once the user provides a list of the game names the algorithm recovers the cluster identities and cluster features of the games. It then applies the algorithm of the content based recommender on the subset of the data with matching cluster identities. The current algorithm takes as an input a list of game names and outputs the 40 most similar games. It also returns a description of the features of the games provided by the user and a visual representation of the similarity of the suggested games. The feature output is useful as it helps the user to understand his taste in board games and if one wishes one can go beyond his comfort zone and explore other game genres. The user also has the option to choose whether he wants to include expansions in the search results and which of the clustering algorithm he wants to use. Because using enhanced clustering segregates the clusters farther apart, it focuses on a smaller subset of the data and can output more recent games than that of the simple clustering. In this sense, the enhanced clustering allows a notion of incorporating the current metagame into the machine. The user can also allow incorporation of average user rating in the algorithm which uses it as a weight vector to multiply the similarity matrix. This weighting clearly favors more recent and popular games as it has been noted that the more recent games have a better user rating in BGG database.

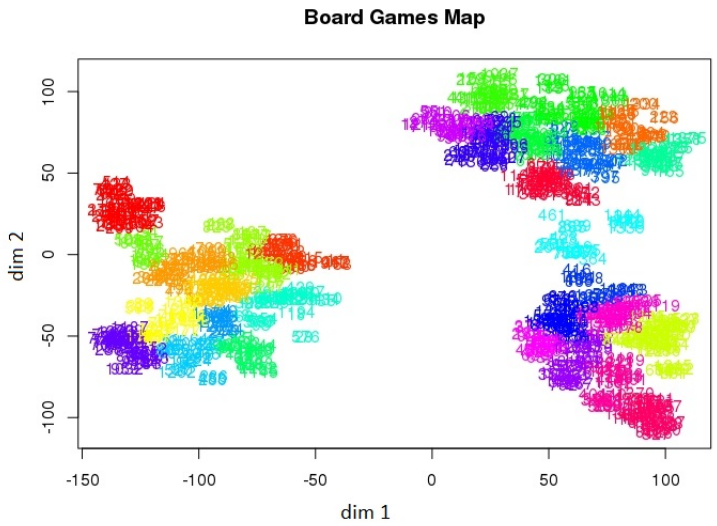
7 Results

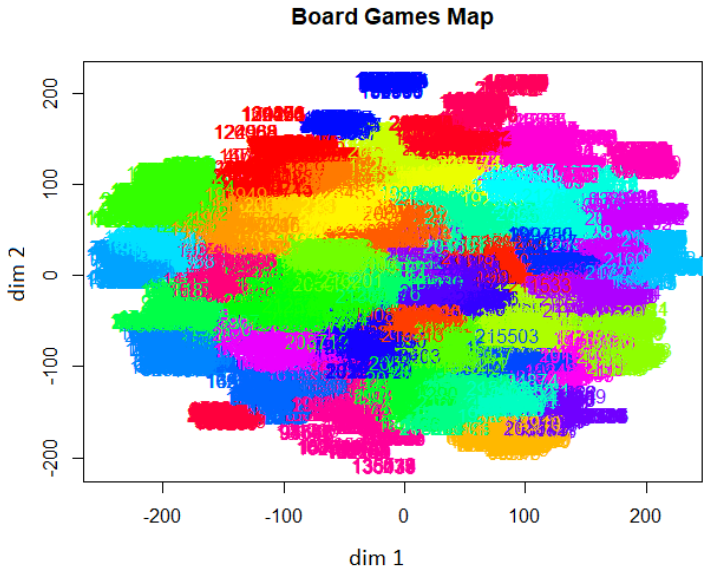
By pre-computing, the cluster assignment of the data by both simple and enhanced clustering algorithm the algorithm becomes very fast. A selection of 6 game titles returns the top 40 similar games along with category information within 5 seconds. The following plots demonstrate the output of the boardgame recommender system. The white node at the center depicts the selection of the user, the details of which are provided at the bottom. The other nodes of the graph are recommendation generated by the algorithm. The distance of the nodes from the center node is based on the similarity score i.e. the closer the nodes are the similar they are. The nodes of the graph are also color-coded according to their cluster assignment and the chart at the bottom right gives information about the genre of the games in the clusters.





The left graph is the recommendation for a user with query words ”Catan”,”Fossil,”Ra” and ”Tikal” whereas the right one is for someone like me with query words ”7 wonders”, ”Scythe”, ”The Resistance”, ”Karmaka”, ”Splendor” and ”Dead of Winter”, some of my favorite board games. Both the graphs were obtained using simple clustering on the full dataset by excluding expansions and allowing user rating. The recommendations provided in the second graph contains games like ”Pandemic” and ”Internal Affairs” which I have already played and liked quite a bit. I am looking forward to exploring the rest. The game genres are mostly correct but can be made more precise.
I also obtained a two-dimensional representation of the board game dataset using the tsne algorithm. The different clusters are color-coded in each picture. The left cluster was generated by using the enhanced clustering on a random subsample of 5000 datapoints whereas the right one was generated by simple clustering on the entire dataset.





Because we used the tsne to obtain a two-dimensional representation the distance between the clusters does not mean much. However, there are clearly 3 major clusters in the left one. This might be due to the fact that the extra weight assigned by the player age, number of players and playing time variables on the distance matrix helps to segregate the clusters more prominently. The meaning of this clustering is explained in the metadata based clustering section. This effect is masked when we use simple clustering to generate the right plot. All the clusters in the left plot have many subclusters. These subclusters denote the effective gameplay mechanic and categories within that specific game genre

and should be the focus of game artists and designers to develop even better games. I for one is looking forward to it.

8 Future Work

The main objectives of the future work are to optimize our clustering algorithm for faster and better prediction. There is no way to check the accuracy of clustering using tsne and maybe another algorithm would have been more superior in that regard. We ignored most of the rating and weight data which seems like a total loss. There is no denying that these contain a wealth of time series information. A thorough exploration is necessary to figure out ways to incorporate these variables into the model. We also threw away a large chunk of the data to avoid user bias arising from user rating. However one can incorporate a rating system like that IMDB's weighted rating formula and cross-validate it using generalized linear modeling. This simultaneously removes the user bias and provides us with a richer data set to begin with.