Textural Analysis on Board Games
Correlation between Game Rules and Game Popularity

Junqian Zhang

University of Milan
Data Science and Economics

Abstract. This project aims at exploring the relation between board game rule setting and game popularity in the board game market. Through the textual analysis, this work tries to find the main aspects about how the game rules are designed and detect the differences among the games on these aspects with text summarization techniques. It employs Canonical Correlation Analysis to detect whether game design can influence the ratings for evaluation on game.

Keywords: Text Summarization · Query Expansion · Canonical Correlation Analysis · Board Game.

1 Introduction

Board Game is one the most popular hobbies in the world nowadays and its market grows rapidly every year. The main idea of this project is to explore whether the game rule design for games has relations with game popularity in the market. The hypothesis behind is that main factors depicted by game designer have been shared by board games in same group, and therefore to study the influence of game rules on the game popularity, we can make use of information retrieval tools to figure out these factors and analyze the correlation between these factors and descriptive variables which evaluate market popularity.

Information Retrieval and Text Summarization develop fast by the contributions of researchers. Great many approaches and algorithms have been proposed. This work mainly inspired by Rahman and Borah 's work [1]. They firstly executed the query-based text summarization and then implemented Canonical Correlation Analysis to reach the final target.

2 Problem Statement and Methodology

The main two questions here are to figure out the main factors in game designing and connect these factors to market metrics. Based on the hypothesis, the games share the common features, and thus aspect-based analysis is adopted. And these aspects extracted can be used as queries for extractive summary. And though the comparison between the summaries, we can test the correlation between the

game aspects and popularity variables and thus evaluate this correlation with ranking variable. More details are provided in the following subsectors.

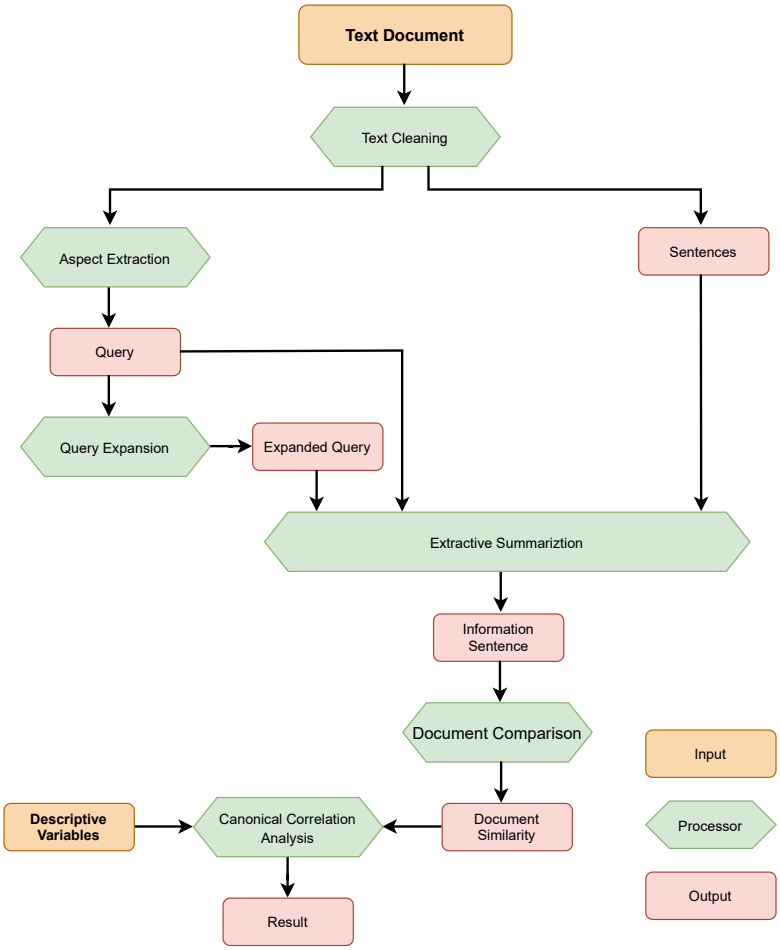

Fig. 1: Project Design

2.1 Text Cleaning

The text analyzed in this project is descriptions of each game as general game rules. For the following steps, all the description are cleaned based on sentence-level. All the sentences are tokenized and words in the sentences are executed lemmatization to get uni-grams by **Part-of-Speech** Tagging, but, at the same time, stop words are not abandoned.

2.2 Query Extraction

Under the assumption that the game rule is composed of settings on different aspects. This target of this step is to find these aspects. These aspects are also called *queries* here because they are used as key words for retrieving the related information in the text. This project focus only on nouns and noun phrases appearing in the text document. All the noun phrases are assumed to be composed of two words, i.e., all the noun phrase are bi-grams.

Based on the uni-grams, bi-grams are built at sentence-level to find candidate noun phrases. This project assumes that the candidate noun phrases should be in the pattern of two nouns, or stop word plus noun or adjective plus noun. So, to efficiently get the bi-grams, bi-grams without nouns inside are abandoned. To obtain the valid noun phrases and rank them, I calculate Pointwise Mutual Information (**PMI**) scores for all the bi-grams. PMI values take into account of the correlation between the two words inside the noun phrase, avoiding the case where the noun phrases in fact are the aspects of the aspects. 50 bi-grams with highest PMI scores are selected but the meaningless ones among them are abandoned.

Except these selected noun phrases, all the other bi-grams are splitted into uni-grams again and corpus is re-cleaned the corpus by removing the stop words. With description as documents, I calculate the term frequency–inverse document frequency (**TF-IDF**) for all the nouns and candidate noun phrases. 20 features with highest TF-IDF values are selected as aspects.

2.3 Query Expansion

To improve the information retrieval performance in later summarization, it is necessary to do query expansion. During the information retrieval, the main task is to figure out all the sentences under same game rule aspect. So synonyms and semantically related words or phrases should be used together with queries found in the last step.

Synonyms and related expressions are connected with queries on the basis of context. Thus, **Word2Vec model** is trained to represent all the uni-grams and bi-grams as vectors, and consequently it is possible to compare the context between each other. These vectors can be seen as a description of the context of each element in the vocabulary. **Skip-gram model** is used to get the input for the word embedding representations to describe the context. Through skip-gram, all the sentences in all the comments are organized into sequences. The length of context window is 3 and each gram concerns the 2 grams surrounded as neighbours.

Since all the words and phrases are in a numeric form, the cosine similarity distance between them can be calculated to see the difference of their contexts.

With the word embedding model, for each aspect extracted in the last step, I take the top 10 words or phrases that are closest to it, which are regarded as different aspect expressions. But some expression may be not nouns or noun phrases, so these are removed from the query set.

2.4 Extractive Summarization

For each description of game, I want to summarize all the *information sentences* which are related to query sets by extractive approach. To achieve extractive summarization, *semantic relatedness score* is built. This index measures how one sentence semantically related to a query set. The higher the index, the higher relativeness between the sentence and the query. As shown in the equation (1), Semantic relatedness score between one sentence S_i and one query Q_m is calculated by summing cosine similarity CS_{iwm} between each word w in sentence S_i and Q_m and dividing with the number of words in the sentence $len(S_i)$.

$$SemanticRelatednessScore(S_i, Q_m) = \frac{\sum_{w}^{W} CS_{iwm}}{len(S_i)} \tag{1}$$

All the expanded queries found in the step 2.3 are replaced by the corresponding query in the sentence. Subsequently, skip-gram model and Word2Vec model are retrained based on new sentences for better context detecting. By calculating the cosine similarities with new models, we are able to get the relatedness scores for all the sentences under each query. For each query set, I extract sentences with related score higher than **0.13** as information sentences.

2.5 Document Comparison

Taking information sentences under one query as a document, I want to compare the similarity between documents to explore the difference on game rule settings. Under each aspect, the documents are represented in numeric vectors by **TF-IDF** approach.

To generate new variables which can represent the meaning behind the game rule setting, the hottest game *Gloomhaven* is set as *reference*, and the similarity between each game and *Gloomhaven* on each aspect is transferred to the new variables. If there is no related sentence in one game, the similarity is set to **0**. In this way, new variables with respect to the game rule aspects can represent the behaviors of games on the aspects.

2.6 Canonical Correlation Analysis

As described in Wang's work [2], **Canonical Correlation Analysis** (CCA) is particularly useful when describing observations that bridge two levels of observation with the aim of modality fusion. This project employs CCA as the tool

to explore the relation between game rule setting and game popularity. We now have two sets of variables from the same data samples. The first set of data contains the variables which can depict the differences of games in rule setting in the previous steps. The second set of data is about the evaluation metrics of game popularity.

By doing CCA, we can identify the canonical variates that are highly correlated to the unknown latent variables behind two datasets. I assume that the latent variable behind is game popularity. Canonical variates quantifies the linear correspondence between the two variable sets based on Pearson's correlation between their canonical variates, so canonical correlation can be used as metrics for successful joint information reduction between two variable sets. In other words, canonical correlation can be used to detect whether any aspects have strong influence on game popularity.

3 Experimental Result

3.1 Data Description

The data analyzed in this project come from the hottest games list on August 28th, 2019 on BoardGameGeek (BGG). The games chosen as target for experiment are those with at least one of tags: Adventure, Exploation, Fantasy, Fighting, Miniatures, Action Queue, Action Retrieval, Campaign / Battle Card Driven, Card Play Conflict Resolution, since they are more possible to be in same group of games and thus it is more reasonable to consider them share the same rule setting aspects. After filtering, there are **4,605** samples, among which *Gloomhaven* is the most popular game and is set as *reference*.

The **7** descriptive variables taken as game popularity metrics are:

- `Average Rating`
- `Standard Deviation of Rating`
- `Average Weight` describes how complexity of a game
- `Own` describes the number of people in possession of the game
- `Trading` describes the number of people willing to trade the game
- `Wanting` describes the number of people who want the game
- `Wishing` is the number of wish lists containing the game

The text regarded as game rules for textual analysis is the `Description` of each game.

3.2 Results and Evaluation

After text cleaning, query extraction and query expansion on variable `Description`, the aspects share by game rules and corresponding expanded queries are listed

in **Table 1**. Since the first variable set takes the first game *Gloomheaven* as reference, the second set of descriptive variables should have similar operations. All the rows of second variable arrays subtract the corresponding value of *Gloomheaven*. After this, both variable sets remove the row of reference.

Query	**Expanded Query**
Action	movement, account, perform, execute, impulse, activation, extent, planning, program, manoeuvre
Battle	combat, confrontation, duel, gunfight, clash, fight, malifaux, conflict, dogfight
Board	flick, hexagonal, game board, interlock, overlay, modular, blank, tall, disc, square
Card	drawn, played, pair, refill, letter, icon, discard, impulse
Character	hero, wrestler, trait, fighter, attribute, battleboard, class, signature, roster
End	end of, game end, continue, victor, feed, remainder, exhaust, vps, finish
Game	gameplay, format, layer, soundtrack, humorous, cd, meaning, the game, this game, meaning, proceeds, battleboard, additionally, fifth, exhibition
Hero	adventurer, character, villain, roster, henchman, survivor, minion, gladiator, posse, warrior
Opponent	your opponent, outwit, opposition, thwart, outmaneuver, authority, attacker, jojo
Play	played, draft, recommend, multiplayer, dealt, reshuffle, accommodate
Player	loser, clockwise, jojo, coach, gangster, challenger, process, proceed
Point	vp, vps, bonus, victory point, prestige, majority, tally
Quot	bang, stuff, loop, sturdy, yes, cabo, erase, undo, write
Rule	beginner, rulebook, ruleset, explain, comprehensive, format, diceland, hardback, introduction
Set	pack, infinity, warhammer, battleground, content, jedi, package
Turn	clockwise, each turn, meter, cell, month, proceed, repeat, forfeit
Use	allocate, combine, drafting, regulate, activate, assortment, interact, manipulation
War	empires, pacific, confrontation, revolution, europe, george, savage
World	deeply, reshape, horizon, the world, realm, greece, universe, history, reality

Table 1: Aspects of Game Rules

After CCA, I firstly look into the first pair of canonical variates. **Fig.2** shows the correlation between them. The canonical correlation between them is **0.39**, and manifestly it is not high. Therefore, the orange regression line in the plot is flat. The color of the points represents their rank in the game list. With different color,

we can clearly observe that, samples with higher rank which means higher popularity are not significantly apart from samples with lower rank. Also, in **Fig.3**, the correlation between the two variates and rank game rank is low. which is **0.2**

But the good thing is that less popular games tend to have lower values in both variates, while more popular have higher values in both. If we level the game ranking into 4 degrees, where each degree has the same amount of samples, as illustrated in **Fig.4**, games of lowest level are more related to two variates.

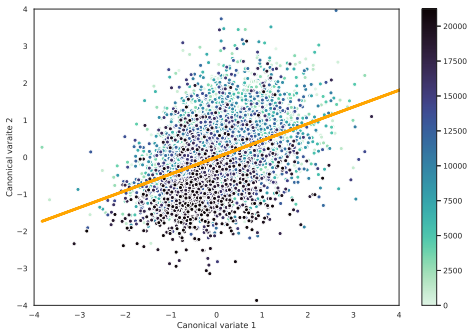

Fig. 2: First Pair of Canonical Covariates colored by Rank

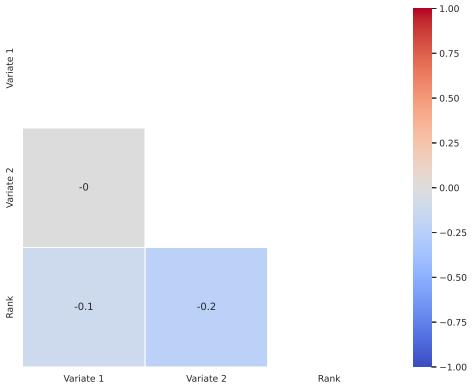

Fig. 3: Correlation Heatmap between Canonical Variates and Game Rank

4 Conclusion Remarks

So far, the result of the project design is not good. On the basis of the game aspects extracted, the correlation between game rule setting and game popu-

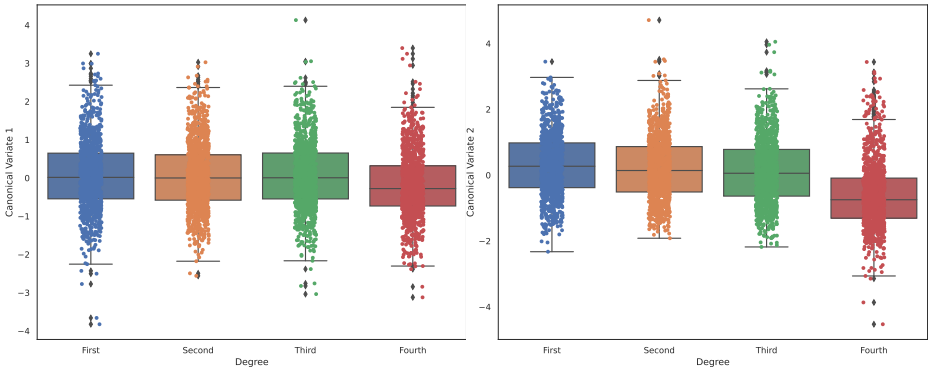

Fig. 4: Boxplot of Canonical Variates and Rank Degree

larity cannot be proved high. But the games with lower ranking supports the relatedness.

Although this project fails to find the relation between game rule settings and game popularity, it does not mean this relation does not exist, and neither does it mean the total design is wrong. This topic deserves more efforts. Firstly, description as game rules is restricted and its content is always general. Its text length is not long enough and thus it cannot provide enough information on rule aspects. If we can replace description by the detailed game rules written by game designers or players, it could provide enough information. Secondly, to compare different documents, I use the top game as reference since the extracted summary is short, usually containing 2 or 3 sentences. If we can more sentences under one topic, it is possible to use abstractive summary to gain a more aggregated summary and in the meanwhile, the comparison metrics can use sub-aspects by topic models.

References

1. Rahman, N., Borah, B. "Improvement of query-based text summarization using word sense disambiguation". *Complex Intell. Syst.* 6, 75–85 (2020)
2. Hao-Ting Wang, Jonathan Smallwood1, Janaina Mourao-Miranda, Cedric Huchuan Xia, Theodore D. Satterthwaite, Danielle S. Bassett, Danilo Bzdok, "Finding the needle in high-dimensional haystack: A tutorial on canonical correlation analysis"