Topic Analysis on User Generated Comments of Chinese Mainland Films

Anyuan Zhong[1(✉)] and Ruiyu Qiu[2]

[1] Shandong Jiaotong University, Weihai, China
223013@sdjtu.edu.cn
[2] The Second Municipal Hospital of Weihai, Weihai, China

Abstract. Online film commentary, which is an important channel for movie fans to express their opinions and emotions, is extremely useful for movie research. In this study, we used topic mining on Douban movie comments to examine 12 popular Chinese mainland films. To provide this, we:1) construct movie comments corpus though web crawling and scraping;2) identify 12 movie topics by word2vec and hierarchical clustering;3) calculate the involved topics of each comment from 12 popular Chinese mainland films;4) evaluate the degree of popularity and praise of each topic of the 12 films. In the result, we give a brief analysis of 12 film topics and choose four of the most essential movie topics: plot, actor, character, and atmosphere, and conduct a visual movie-topic analysis based on two factors: popularity and praise.

Keywords: film analysis · topic mining · hierarchical clustering · word2vec

1 Introduction

Because of its convenience and low threshold, online film commentary has grown in popularity among film fans. Online film comments come from a diverse variety of moviegoers and provide a wealth of vital information on film opinions, feelings, and emotions. As a result, data mining of film comments has become a crucial tool for film research. Ramos et al. analyzed independent film comments posted on Twitter through keyword analysis and topic analysis [1]. Sun and Gai analyzed the film reviews of the Chinese animated film "Ne Zha" by combing emotional analysis and the LDA model [2]. Xue et al. collected 3000 Douban comments and conducted sentiment classification with the naive Bayes algorithm [3]. Quan et al. extracted the keywords of the reviews of the micro movie "What is Peppa" with word2vec and network analysis aiming to illustrate the movie's core theme [4]. Wu et al. explored the typical opinion of Douban comments on the animated film "Monkey King: Hero Is Back" through TFIDF keywords extraction, word2vec, and K-means clustering [5]. In general, the above research can be summarized into the following two categories:

- Content analysis. The approaches always include keyword analysis, topic analysis, visualization, and other techniques, and this area deals with providing intuitive, systematic, and thorough descriptions of a movie.

M. F. b. Sedon et al. (Eds.): SSHA 2023, ASSEHR 752, pp. 379–386, 2023.
https://doi.org/10.2991/978-2-38476-062-6_48

- Sentiment analysis. This category concerns identifying the sentiment polarity or classification of film comments, which is always done in conjunction with topic mining to gain a thorough understanding of the feelings of viewers.

The major relevant research focused on the analysis of a single movie or several movies, and the comment data supporting the analysis is limited. This article aims to give a comprehensive and systematic knowledge of Chinese mainland movies with text mining on huge amounts of user comments and reviews of a wide range of films.

2 Method

Our process consists of three steps. First, we construct a database of 916 movies, including their profile, short comments and review articles. Second, we compute the word vector with all the short comments and review articles, and identify the major movie topics with clustering. Finally, we filter the popular Chinese mainland movies, quantize the topic popularity and praise based on the movie's short comments, and conduct the movie-topic analysis. The framework of our research is shown in Fig. 1.

2.1 Data Acquisition and Preprocessing

We developed a web crawler to collect the original data with python (version 3.7.9) and selenium (a web automation tool). The web crawler visited movie pages on Douban, loaded the comments and reviews automatically, and saved the web pages in HTML format on disk. Then, we extracted the following data about each movie using beautifulsoup:

- Basic data, including the name, score, release time of each movie.
- Short comments, including the content, movie rating and support num of each short comment.
- Review articles, including the content and support num of each review article.

Finally, we conducted word segmentation with Jieba on the content of short comments and review articles, turning the textual data composing of sentences to word lists.

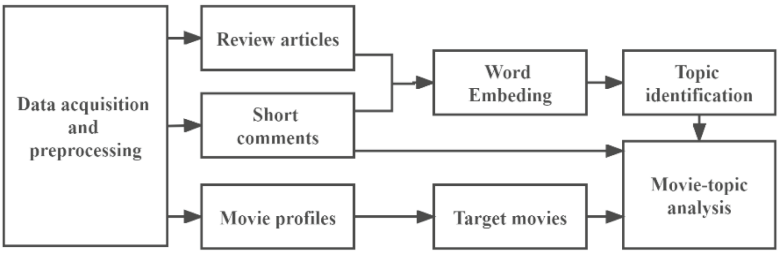

Fig. 1. The research framework.

We chose Douban's high score list, hot list, latest list and Chinese language movie list as our data sources, and build a movie database including 916 movies. The database contains 439,137 short comments and 351,672 review articles, totaling 286,286,268 words after word segmentation.

2.2 Word Vector Model

Word2Vec is a Google open-source word NLP tool that can provide word representation in vector space. Word2vec employs a neural network to map discrete words to low-dimensional vector space based on their position in the training set [6]. Each word has a semantically relevant vector in a word2vec output, and the similarity of two word-vectors can properly reflect the semantic similarity between the two words.

To train the word vectors, we utilized the Word2Vec model from the Python Gensim package. The CBOW method predicts the central word based on its context, whereas the SkipGram method predicts the context based on the central word. The SkipGram approach is more competitive for low-frequency words, but it takes longer. We finally picked the CBOW method to train the word vector model since low-frequency words cannot be used as feature words for later topic identification. In our code, the vector size was set to 128, and all other parameters were left at their default values. We used all of the short comments and review articles to increase accuracy.

2.3 Topic Identification

Because words with similar vectors have comparable meanings in the corpus, a vector-based clustering algorithm can be used to group words with high semantic similarity to the same cluster, and each cluster can be used as a candidate topic. The subject identification process consists of two steps: 1) Feature extraction to reduce the number of words. 2) Hierarchical clustering on the feature words to identify the candidate topics.

2.3.1 Feature Extraction

TFIDF and TextRank are two feature extraction algorithms that are extensively utilized in NLP. The TFIDF method identifies keywords based on the product of TF (Term frequency) and IDF for each word (Inverse document frequency), which may drop the high-frequency word in the corpus and ignores the relationship between words. TextRank algorithm uses a graph model-based sorting algorithm to determine keywords. In this algorithm, words are taken as network nodes and the co-occurrence frequency between words is used to determine the weight of edges between nodes. Before the algorithm starts, a fixed TextRank value is set for each node, and then the TextRank value of each node is iteratively updated by a weight calculation formula until the TextRank value converges. Compared with the TFIDF algorithm, the TextRank algorithm makes full use of the relationship between words and pays more attention to the distribution of words.

The goal of this sector is to extract keywords that are indicative of the complete movie database. Given this goal, we believe that the TextRank algorithm is better appropriate for our research. To extract the feature words, we used the Jieba package's TextRank function. First, we excluded from the database all words other than nouns, verbs, and

adjectives. Second, for each movie, we use the words with the highest 20% TextRank value as the feature words. Finally, we combine all of the movie's feature words as the database's feature words.

2.3.2 Hierarchical Clustering

This section seeks to identify candidate subjects by clustering on feature words. Because of its tree-like clustering process, hierarchical clustering was chosen as our clustering method. During the clustering process, one can analyze the cohesiveness of each node and select the solution with appropriate inner-cluster semantic similarities. The process is as follows:

1) Generate the initial clusters set by regarding each feature word as a cluster.
2) Compute the inter-similarity of each two clusters with formula 1, and merge two clusters with the highest inter-similarity:

$$\mathrm{sim}(c1, c2) = \underset{w1 \in c1, w2 \in c2}{\mathrm{average}} (\mathrm{cos\_sim}(v_{w1}, v_{w2})) \tag{1}$$

where c1 and c2 are two clusters, wi is a feature word of ci, Vw is the word vector of w, and cos_sim refers to the cosine similarity of two vectors.

3) Repeat 2) until only one cluster remains.

2.4 Movie-Topic Analysis

The topics excavated from massive movie comments and review articles are well reflective of moviegoers' concerns. This section seeks to provide an approach to evaluate the popularity and praise of each topic in the context of given movies, which is supposed to be beneficial for the policy-making of the movie industry. The short comments on Douban are featured by high information density. Furthermore, each short comment is linked with the author's movie rating, which reflects the author's attitude toward the film. We used the short comment texts and the combined movie rating as the foundation for movie-topic analysis for the reasons stated above.

First, we identified the relevant topics in each short comment. We assumed a brief comment with word list S and a topic with feature word set L are related if one of the following criteria is met: 1) S and L intersect. 2) There is at least one word in S that has a cosine similarity to all words in L that is bigger than 0.6. Second, we evaluated each topic's popularity and appraise of a given movie by formula 2:

$$\begin{aligned} \mathrm{Populairty(t)} &= \sum_{i} C_{it}/N \\ \mathrm{Praise(t)} &= \sum_{i} S_i C_{it}/N \end{aligned} \tag{2}$$

where t refers to a topic, N refers to the number of short comments, The value C_{it} is 1 if the ith comment is related to topic t and 0 otherwise, The value of s_i indicate the movie rating linked the ith short comments (1 represents Five stars, 0.75 represents four stars, 0.5 represents 3 stars, 0.25 represents two stars, and 0 represents zero stars).

3 Results

In this section, we present the results we found for Chinese mainland popular movies released after the year 2015. We choose 12 movies as a case study base on the criteria that the number of comments is greater than 700,000 and the rating is between 7 and 8. The selected films are shown in Table 1.

3.1 Overview of Topics

By topic mining in the last chapter, a total of 12 movie topics are obtained, which are the character, implication, plot, lines, atmosphere, OP&ED, touching, actor, structure, tensity, lens and music. Their feature words are shown in Table 2.

The popularity and praise of the above topics in the 12 domestic movies, together with their average values, are shown in Fig. 2. It can be seen from the figure that the most popular movie topics are plot, actor, atmosphere and character. The praise value of these four themes is around the average, and the praise value of actors and atmosphere is slightly higher than that of plot and character. Plot and characters are two important aspects of movies, and they complement each other. Chinese domestic movies need to further improve in these two aspects. The four topics of structure, OP&ED, touching and lens are in the middle, and the praises of touching and OP&ED are significantly higher than the average, indicating that some films have outstanding performances in these two aspects. The remaining four least popular topics are lines, tensity, implication and music. The praise of music and tensity is significantly higher than the average, and it is necessary to explore the shining spots worthy of further development.

Table 1. Representative Chinese mainland movies.

Movie	Year	Average Rating	Comments number
Operation Mekong	2016	8	703853
The Wandering Earth	2019	7.9	1784160
Mo seung	2018	8	884085
The Eight Hundred	2020	7.5	769158
Youth	2017	7.7	765100
The Island	2018	7.1	856733
A Cool Fish	2018	8	969413
Visual	2019	7.5	835654
White Snake	2019	7.8	709692
My People, My Country	2019	7.6	957755
Hi, Mom	2021	7.7	1306478
A Little Red Flower	2020	7.2	712254

Table 2. Feature word of each topic

Topic	Feature Words
Character	Figure, Role, Protagonist, Hero, Heroine, Image
Implication	Metaphor, Symbol, Mean, Fable, Image, The moral
Plot	Story, Plot, Scene, Detail, Part, Scenario
Lines	Lines, Dialogue, Narrator, Communication, Monologue
Atmosphere	Rhythm, Scene, Spectacle, Atmosphere, Special effects
OP&ED	Start, Ending, Beginning, Opening, Finishing
Touching	Shock, Touching, Moving, Amazing, Impressing
Actor	Actor, Acting skill, Performance, Starring, Supporting role
Structure	Setting, Process, Design, Reverse, Orgasm, Matting
Tensity	Fearing, Exciting, Nervous, Compact, Stimulus
Lens	Lens, Picture, Color, Long lens, Close-up, Light and shadow
Music	Music, Ringing, Song, Theme song, Dancing

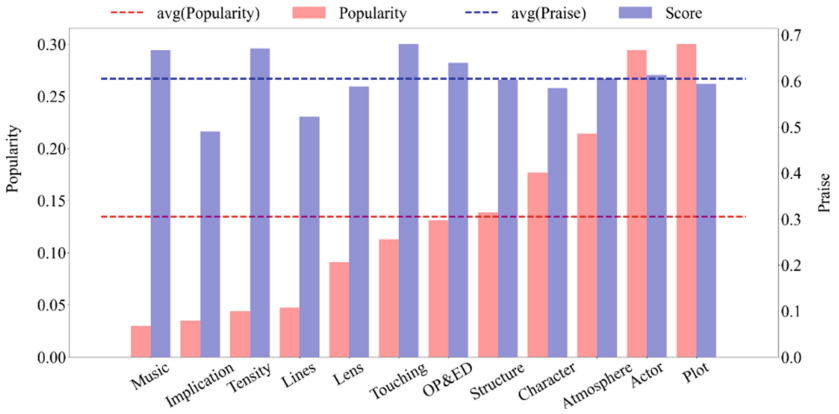

Fig. 2. The popularity and praise of each topic

3.2 Topic Based Film Analysis

Different movies have their different characteristics. In this section, four topics with the highest popularity were selected to analyze the popularity and praise of each movie. We plotted a scatter plot for each topic, where the horizontal and vertical coordinates are the popularity and praise ratings of the film on that topic, and the average value is shown as a dotted line. As can be seen from Fig. 3, for each topic, the two mean lines can divide the movies into four parts.

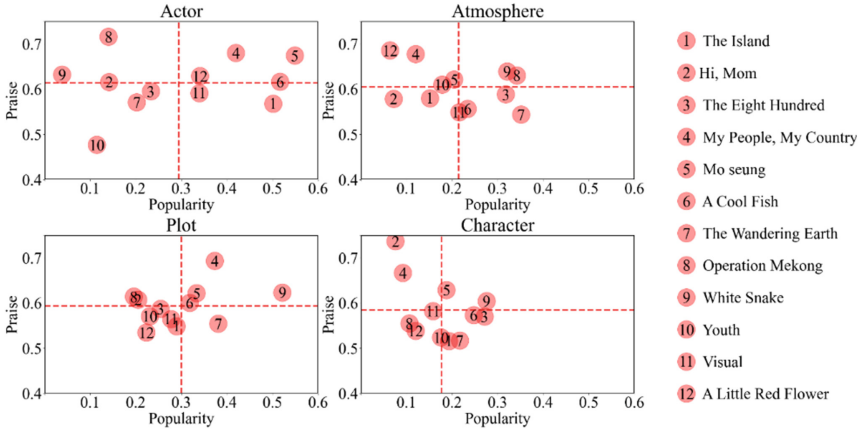

Fig. 3. The scatter plot of the 4 main topics

- The topic right area: The movie has high popularity and praise on this topic, which is the highlight of the film, such as the actor topic and plot topic of "Mo Seung", the atmosphere topic of "White Snake", and the plot topic of "My People My Country".
- The bottom right area: The popularity of the film on this topic is high, but the praise is low, which is easy to have a negative impact on the word of mouth of the movie. Lessons need to be learned to improve. For Example, in the actor topic of 'The Island', the performance is exaggerated, causing the viewers aversion.
- The top left area: The topic is not popular in the film, but it is highly praised. It indicates that the film has a good performance in a certain aspect, but the attention is relatively low, which can be improved by means of publicity and marketing. For example, "Operation Mekong" is an action film, the actors receive less attention, but the leading performances are praised.
- The bottom left area: The popularity and praise of the topic of the film are low, such as the actor topic of "Youth".

4 Conclusion

In this paper, we used a large number of user generated comments and reviews to mine the topics of Chinese mainland movies, and conducted visual analysis. On the whole, the four topics of plot, actor, atmosphere, and character of Chinese mainland films have attracted the most attention, and the topics of touching, tensity, music, and OP&ED have been highly praised. For the four most popular topics, this paper visually analysis from two dimensions of popularity and praise in the context of 12 Chinese mainland movies. Our research is meaningful for understanding Chinese mainland movies comprehensively, which is supposed to be conducive to the development of China's film industry. As a limitation, the calculation of praise is based on comment ratings, we found in some comments, there is a gap between the real feeling reflected by the content and the rating. In some papers, sentiment analysis is used to evaluate the feelings contained in a comment, but the error is large because of the diversity of expression. In further research, we will focus on improving reliable method to evaluate the moviegoers' feelings.

References

1. Ramos, C.D., M.T. Suarez, and E. Tighe, Analyzing National Film Based on Social Media Tweets Input Using Topic Modelling and Data Mining Approach, in Computational Science and Technology. 2019. p. 379–389.
2. Sun, S., et al., Research on User Comments of Douban Animation Made in China Based on Text Mining Technology, in Proceedings of the 2019 7th International Conference on Information Technology: IoT and Smart City. 2019. p. 89–93.
3. Xue, J., et al. Analysis of Chinese Comments on Douban Based on Naive Bayes. in Proceedings of the 2nd International Conference on Big Data Technologies. 2019.
4. Hamid, N., et al., Keyword extraction for film reviews based on social network analysis and natural language technology. E3S Web of Conferences, 2020. **189**.
5. Wu, T., F. Hao, and M. Kim, Typical opinions mining based on Douban film comments in animated movies. Entertainment Computing, 2021. **36**.
6. Mikolov, T., et al., Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.