Keyword extraction for film reviews based on social network analysis and natural language technology

Quan Yanan[1,*], Tan Fuqiang[2]

[1] Jiangxi University of Science and Technology, Ganzhou, Jiangxi, 341000, China
[2] Shenzhen University, Shenzhen, Guangdong, 518000, China

Abstract. At present, there are many movie reviews appear on main stream websites, and these evaluations are quite different to the same movie. As a customer, how to choose your favorite movie and television program? To solve this problem, this study attempts to use the semantic analysis of word vectors (Word2vec) semantic analysis in machine learning as a research tool to mine a large number of movie reviews. The research shows that most movie reviews have a certain theme cohesion and their semantic network has quite connected. Through the use of social network analysis and the use of Word2vec word vector technology in natural language processing, it is possible to present a streamlined movie review based on movie review network semantics and keyword extraction, thus helping to select the favorite movie review.

1 Introduction

Movie review, or film review for short, it means the analysis and comment on the director, actors, lenses, photography, plot, clues, environment, colors, lighting, audio-visual language, prop functions, transitions, and edits of a movie. By analyzing, appraising and evaluating the aesthetic value, cognitive value, social significance, lens language that displayed in the screen, the purpose of movie review is to achieve the goal of shooting the film, to explain the theme expressed in the film, moreover, by analyzing the success or failure, the gains and losses of the film, it helps the directors to broaden their horizons and improve their creative level, so as to promote the prosperity and development of film art; at the same time, they can affect the audience's understanding and appreciation of the film through analysis and evaluation, and improve the audience's appreciation level, thereby indirectly promoting the development of film art. Therefore, as a condensed description of the comprehensive performance of a movie, film criticism is self-evident for the audience's choice of viewing on a movie screen. Previous studies on film criticism have mostly used film criticism content as qualitative analysis research materials, and then tapped the essence of related film criticism. For example, MM Lauzen, DM Dozier (1999) incorporated film reviews into the retrospective literature of film and television research, thereby enhancing the link between film and television and their reviews. E Lindasari, K Ansari, M Marice (2019) applied interactive multimedia technology to the study of high school students' film reviews. AF Alsaqer, S Sasi (2017) and others used rapid miner to study movie reviews. A Tripathi, SK Trivedi used multi-feature selection techniques to analyze the reviews of Indian movies. A Khan, MA Gul (2020) studied movie reviews using a sorting algorithm based on supervised learning and graphs. SS Sharma, G Dutta (2018) judged the polarity of movie reviews. GS Brar, APA Sharma (2018) used supervised machine learning models in machine learning to study the sentiment of movie reviews. H Park, K Kim (2019) used the CNN-LSTM model to classify the sentiment of movie reviews. F WU, S LI, G ZHOU (2019) used the LSTM model to classify the content of movie reviews.

The previous research only carried out pure artificial intelligence extraction of movie reviews, and ignored the semantic relationship between movie reviews. In view of this, this study uses Social Network Analysis (SNA) combined with Word2vec word vector technology in natural language processing, which is intended to explore the internal relationships and keywords of movie reviews to help the audience in selecting their favorite movies.

2 Introduction of research methods

2.1 Research data

This article takes the movie review of the micro movie "What is Peppa " as the original data. According to the research needs, first of all, it collects a total of 30 related movie review articles including "What is Peppa " at the level of topics, keywords and research topics; Secondly, collected a total of 3 academic research articles of this film.

* Corresponding author: 278001226@qq.com

2.2Research tools

Due to research needs, this study uses Word2Vec (Word to Vector) as a research tool in natural language processing. First of all, the word vector was first proposed by Hinton, and Bengio et al. established the earliest original model of the word vector. This method can be divided into two types: One-hot Representation and Distributed Representation. The former has a simple representation method but limited semantic expression capabilities. The latter is based on the former's advancing model, to a certain extent, it makes up for the former's limited semantic expression capabilities and the sparse and long matrix. Secondly, the Word2vec tool is a natural language processing tool launched by Google in 2013, and the internal algorithm draws on the basic concept of the Neural Network Language Model (NNLM). The advantage is that through a given corpus, the words in the text can be mapped to the real vector space, and the real vector space is composed of multiple dimensions, each of which can represent the corresponding shallow semantic features. Finally, the mature Word2vec tool is mainly divided into CBOW and Skip-Gram models. Due to the large number of training sets studied in this experiment, the Skip-Gram model that is expected to measure the related words by entering a word is expected. The model has the characteristics of precise semantics and excellent performance in large training sets.

2.3 Research process

2.3.1Natural language mining process

The research is divided into four processes: first, the original corpus preprocessing, second, real word extraction, model operation again, and finally the research results are presented. The detailed research process is shown in Figure 1:

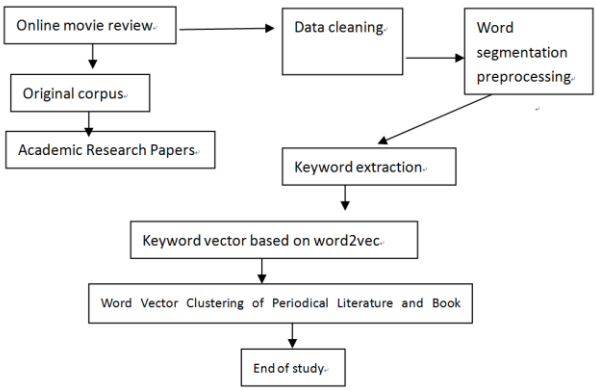

Fig. 1. Research process of movie review mining based on word2vec.

According to the characteristics of online movie reviews and academic research papers, this study uses natural language processing technology to analyze the keywords of these two aspect. The specific process is as follows:

First, the original corpus is preprocessed. Theoretically, the research technique used in this research is applicable to texts in multiple languages, but since this research uses Chinese text, the original corpus needs to be processed before the experiment begins, and the irrelevant stop wordsto this experiment is removed to avoid corresponding noise interference, thereby improving the precision of the topic. In addition, in order to improve the extraction of real words in the later stage, and then label the corresponding parts of speech, the female labeling tool used in this experiment is the NLPIR Chinese word segmentation system.

Second, real word extraction. The keywords of academic research papers are usually given by relevant authors, and have the characteristics of high accuracy, fitting the theme of the article, and the guidelines for streamlining the full text. Therefore, in the past research, keyword extraction was usually used for research. However, this study requires full-text inspection of all data, which is intended to extract nouns, verbs, adjectives and adverbs in the text, so a more classic keyword weight calculation formula (Term Frequency–Inverse Document Frequency, TF-IDF) is required. By calculating the TF-IDF value of the extracted vocabulary to filter the vocabulary whose TF-IDF value is less than the established standard, the corresponding real word set is finally formed. The TF-IDF calculation formula is:

$$Q(t,d_i)=\frac{tf(t,d_i)\times \mathrm{idf}(t)}{\sqrt{\sum_{t\in d_i}\left[tf(t,d_i)\times \mathrm{idf}(t)\right]^2}} \qquad (1)$$

Again, the model runs. One is the acquisition of real word vectors based on Word2Vec technology. To a certain extent, online movie reviews and academic research papers have common attributes due to the focus of the subject, and there is also a clear connection. This is also the basis of this research. However, in fact, after passing the preliminary experiments, it was found that the two are still quite different. This study believes that although there are certain differences between the two, there are cases where the keywords do not correspond, and the same term is expressed differently in online movie reviews and academic research papers. However, because this study has the theme of concise characteristics, the terminology of application environment is also similar, so this problem can be better solved by judging the context information of the vocabulary. The second is to obtain the theme and theme vector representation through the calculation of keyword word vector clustering. This experiment adopts keyword word vector clustering based on X-means. This algorithm is an improved version based on K-means algorithm. Its advantage is that it does not need to formulate the number of clusters K at the beginning of the operation. The range of values is sufficient. Since then, the algorithm will find the final number of clusters K within the specified range by calculation to achieve the optimization of keyword word vector clustering division. The third is to calculate the semantic similarity of subject words based on word vectors. The semantic similarity is determined by the cosine distance between the collection of two real words. In this experiment, the modified angle cosine formula is used to calculate the

semantic similarity of the subject words C1 and C2, assuming that the C1 subject has the following collection {W11, W12,..., W 1m}, and the actual word collection under the C2 subject word is {W21, W22,..., W 2n}, and m>n. The formula for calculating the angle cosine is:

$$Sim(C1,C2)=\frac{\sum_{k=1}^{n} VSim_k \times WC_{2k} \times WC_{1k}}{\sqrt{\sum_{k=1}^{n} WC_{1k}^2 \times \sum_{k=1}^{n} WC_{2k}^2}} \quad (2)$$

Finally, through the above procedures, this study has obtained relevant research results.

2.3.2 Introduction to Social Network Analysis and Exploration of Semantic Network

First of all, the social network analysis method is a quantitative analysis method developed by sociologists based on mathematical methods, graph theory, etc. In recent years, this method has been widely used and played an important role in career mobility, the impact of urbanization on individual happiness, world politics and economic systems, and international trade. Social network analysis is a relatively mature analysis method in the field of sociology. Sociologists can use it to explain some sociological problems easily. Experts in many disciplines such as economics, management, etc. in the new economy era-the era of knowledge economy, when facing many challenges, began to consider borrowing the research methods of other disciplines, social network analysis is one of them. The network refers to various associations, and the social network (Social Network) can be simply referred to as the structure of social relations. The problem of Social Network Analysis (SNA) stems from the adaptive network in physics. By studying network relationships, it helps to combine inter-individual relationships, "micro" networks with the "macro" structure of large-scale social systems. As a result, quantitative analysis methods such as mathematical methods and graph theory are a branch of research that has gradually developed in the fields of sociology, psychology, anthropology, mathematics, and communication science since the 1970s. From the perspective of social network, human interaction in the social environment can be expressed as a pattern or rule based on relationship, and the regular pattern based on this relationship reflects the social structure. The quantitative analysis of this structure is the starting point of the social network analysis. Social network analysis is not only a tool, but also a relational way of thinking, and it can be used to explain some problems in the fields of sociology, economics, management and so on.

Second, the semantic network is a form of expressing human knowledge in a network format. It is one of the representation methods used by artificial intelligence programs. It was proposed by J. R. Quillian in 1968. It was first proposed as an obvious axiomatic model of human associative memory, and then used in AI for natural language understanding to express propositional information. In ES, the semantic network is implemented by PROSPEUTOR, which is used to describe the concept and state of objects and the relationship between them. It is composed of nodes and arcs between nodes. Nodes represent concepts (events, things), and arcs represent the relationship between them. Mathematically, the semantic network is a directed graph, corresponding to logical notation. A semantic network is a structured way to represent knowledge with graphs. In a semantic network, information is expressed as a set of nodes, and the nodes are connected to each other through a set of marked directed lines, which are used to represent the relationship between the nodes. In artificial intelligence programs, predicates and their arguments can be regarded as nodes in the semantic network; the lattice relationship is equivalent to the connection form between nodes. Semantic networks are a semantic-oriented structure. They generally use a set of inference rules, which are specially designed to correctly handle the special arcs that appear in the network.

3 Research results

3.1 Natural language mining results based on word2vec

Using word2vec natural language processing technology to analyze the movie reviews of the movie "What is Peppa", we extracted relevant keywords to indicate what the core theme of the movie is. This allows an audience to understand the theme of the film without anyone telling the content of the film "What is Peppa ". The keywords of the word2vec movie review of "What is Peppa" are shown in Table 1:

Table 1. Keywords of the movie review of "What is Peppa" based on word2vec

sequence	Keywords	Keyword Weight
1	Grandpa	0.01314893338540138
2	Rural	0.008701232995031066
3	Li Yubao	0.006718881258110817
4	New Year	0.006354759124862145
5	grandson	0.006009156776539744
6	son	0.004581335184776074
7	Movie	0.004443473005277963
8	parents	0.0037258942552046335
9	Affection	0.0036802026579924258
10	China	0.003569633994340517
11	Little Pig	0.002916451176226477
12	Smartphone	0.0034042939635141028
13	Chinese New Year	0.0027824245111827783
14	Old people	0.0026806644063599797
15	Movie	0.002598312821165368
16	students	0.0025837647942325757
17	Image	0.0022823447482943826
18	Gift	0.00225664401586477
19	Family	0.0022152323386651833
20	Humanity	0.002184270517772588

It can be seen from the table that the film review of "What is Peppa" shows more faithfully that the film is a short film about the grandfather of the country preparing the Spring Festival gifts for the grandsons in the city. The film emphasizes the relationship between the grandfather and the grandson. I didn't see the cartoon "Peppa " and never used a smartphone, but my grandfather learned the gift needed by grandson's mouth through a voice call with grandson: Peppa . So Grandpa started making it himself. This makes the grandpa's Peppa look strange, but the short film has won praise from the audience because it highlights the family ethical care of the Chinese and the true love between the generations. As far as the above keyword mining is concerned, it can reflect the main spiritual expression of the short film to a certain extent. This shows that the word2vec model is really unique in terms of semantic analysis. However, the keyword selection does not completely explain the relationship between the movie reviews of "What is Peppa". It can only show part of the face of the movie review. Therefore, this Through the use of social network models, the research explores the network semantics between the movie reviews of "What is Peppa ", to show the core expression of the short film, and finally achieves help to prevent the audience from watching "What is Peppa ?" In the case of a short film, learn about the content and emotional theme of the short film.

3.2 Mining results based on Social Network Analysis (SNA)

Research on social network models and semantic network analysis shows that the short film "What is Peppa" presents results of text information mining that are similar to word2vec topic mining.

Among them, at the core position of the network semantic graph are information such as "Peppa", "grandpa", "family", "aging", "new year" and "grandson". As far as this information is concerned, the research can also infer that the significance of the picture is to use the film review as the research material to mine the core tasks of the film "What is Peppa" and the core ideas to be expressed. The specific network semantic vocabulary is shown in Table 2:

Table 2. High-frequency vocabulary of video comment

Sequence	keywords
1	Peppa
2	grandpa
3	rural area
4	have the Spring Festival
5	grandson
6	Chinese New Year
7	Cell phone
8	Old man
9	Short film
10	movie

Through Figure 1 and Table 2, we can more clearly ecognize: First, the film review of "What is Peppa" thinks that the film mainly tells the story between the grandfather and grandson in the Spring Festival periodical. Secondly, the film review of "What is Peppa" thinks that the protagonist of the film is grandpa and grandson; finally, the short film of "What is Peppa" mainly expresses the matter that the rural grandfather prepares a grandson gift for the grandson.

4 Conclusion

This research uses the word2vec word vector semantic analysis tool, social network analysis (SNA) and semantic network analysis tool in natural language processing technology. It studies how to make the audience get the characters, content and express their thoughts through film review, and then provide an auxiliary role for the audience to choose whether to watch the movie. This study proves that the combination of the word2vec word vector semantic analysis tool, Social Network Analysis (SNA) and semantic network (semantic network) analysis tool can more clearly show the content,the themes and many other information of movie reviews through experiments on case exploring.

However, this study also has certain limitations. For example, the use of centralized tools is a bit more complicated. The next step of the research needs to develop corresponding algorithms to synthesize the performance of these three tools.

References

1. Lauzen, M. M., & Dozier, D. M. The Role of Women on Screen and behind the Scenes in the Television and Film Industries: Review of a Program of Research. Journal of communication inquiry, **23(4), 355-373.**(1999).
2. Lindasari, E., Ansari, K., & Marice, M.. Interactive Multimedia Development in Learning of Film Review Text for 8th Grade Students in Senior High School (SMP) 1 Tanjungmorawa. Budapest International Research and Critics in Linguistics and Education (BirLE) Journal, **2(4), 355-362.(**2019)
3. Alsaqer, A. F., & Sasi, S. Movie review **summarization** and sentiment analysis using rapidminer. In 2017 International Conference on Networks & Advances in Computational Technologies (NetACT) **(pp. 329-335).** IEEE.(2017, July).
4. Tripathi, A., & Trivedi, S. K.. Sentiment analyis of Indian movie review with various feature selection techniques. In 2016 IEEE International Conference on Advances in Computer Applications (ICACA) **(pp. 181-185).** IEEE. (2016, October)
5. Khan, A., Gul, M. A., Zareei, M., Biswal, R. R., Zeb, A., Naeem, M., ... & Salim, N. Movie Review Summarization Using Supervised Learning and

Graph-Based Ranking Algorithm. Computational Intelligence and Neuroscience, (2020)

6. Sharma, S. S., & Dutta, G. . Polarity Determination of Movie Reviews: A Systematic Literature Review. International Journal of Innovative Knowledge Concepts**, 6, 12.**(2018)
7. Brar, G. S., & Sharma, A. P. A.. Sentiment Analysis of Movie Review Using Supervised Machine Learning Techniques. International Journal of Applied Engineering Research, **13(16), 12788-12791.** (2018)
8. Park, H. Y., & Kim, K. J. . Sentiment Analysis of Movie Review Using Integrated CNN-LSTM Mode. Journal of Intelligence and Information Systems, **25(4), 141-154.**(2019)