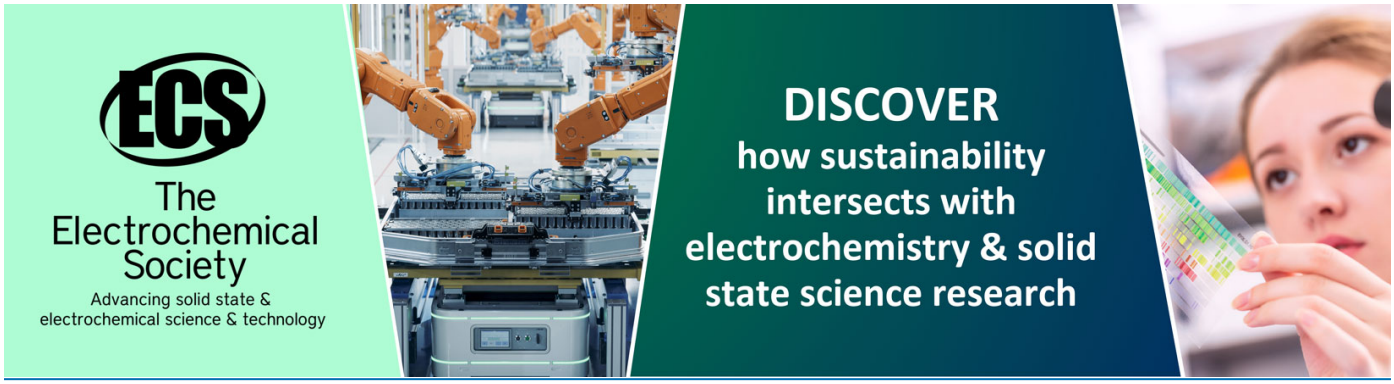
Multi-Angle Movie Reviews Analysis Based on Multi Model



Multi-Angle Movie Reviews Analysis Based on Multi Model

Yanzhe Liu [1], Bingxiang Liu[1,*], Jiajia Yu [2], Zhijian Yu [1]

[1]Jingdezhen Ceramic Institute, Jingdezhen 333403, China

[2]Shaanxi University of Science & Technology, Xian 710021, China

*lbx1966@163.com

Abstract. In this study, movie reviews are used as data sets to extract related phrases, topics, and sentiment scores from the text. Based on users' information, users' behavior preferences and their influences are analysed, and text semantic information is mined from multiple perspectives. A variety of data processing and machine learning methods including text segmentation, Apriori association rule mining algorithm, sentiment analysis, linear fitting, TF-IDF algorithm, PCA dimensionality reduction, and LDA topic model is used in the research. At the same time, due to the coarse granularity of the topic extraction in the LDA algorithm, it is not suitable for short text, this paper proposes a new topic model based on improved k-means and TextRank and gets good results on this dataset. This paper uses multiple data mining models to analyse film reviews and presents an empirical study of the efficacy of machine learning techniques in text semantic mining.

Keywords: Text Mining, Natural Language Processing, Visualization, Topic Model.

1. Introduction

The growing text data in the current network hides a lot of information, these data have considerably exceeded the ability of manual analysis, what is hard nowadays is not to confirm the availability of information but how to extract information in the vast ocean of content. Therefore, it is worth researching automatic processing and analysing text data. This paper uses movie reviews data as an example to explore this topic.

The methods used in previous studies are limited to one of sentiment analysis or abstract extraction [1-5], but in fact, the information contained in movie reviews is rich and diverse, and a single processing scheme will ignore other important information. This research uses a variety of text analyses and machine learning methods, and at the same time, it analyses the comments completely from multiple angles.

2. Methodology

This chapter mainly discusses the data used in this study and the main methods of data analysis. Explained in turn how to preprocess the data, how to find the connection between comments, how to reduce the dimensionality of the data, how to vectorize the text, how to extract comment topics, how to analyse the sentiment value of the text, and how to analyse users' preference for watching movies and the influence of preference on a movie score. The data processing flowchart is shown in Fig.1.

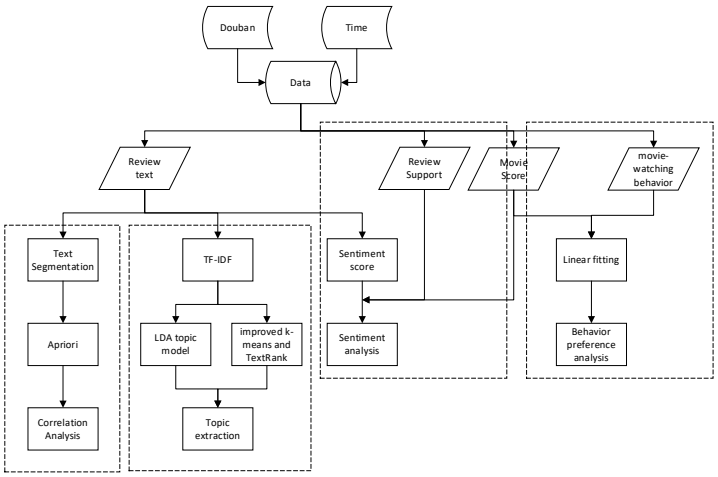


Fig.1 Data processing flowchart

This article uses crawler technology to obtain movie review information and user personal information from two mainstream movie review sites in China, and preprocess the data. The data preprocessing mainly includes synonym replacement, cleanup of useless data, conversion of traditional Chinese to simplified, and text segmentation. Adding exclusive words in the data set to expand the word segmentation dictionary, and using the expanded dictionary to segment the text, and finally using the stop word list of HIT to remove the stop words.

After word segmentation, this paper constructs the result of word segmentation into a vocabulary matrix, uses the Apriori algorithm [6,7] to conduct correlation analysis on the constructed vocabulary matrix, expresses the connection between each comment through strong association rules, and finds common discussion topics and attitudes.

In order to improve the processing speed and accuracy, this paper firstly uses the TF-IDF algorithm to perform feature selection and vectorization of the text before processing [8]. The method is to calculate the TF-IDF value of each word, and finally expresses the text as a vector of weighted terms appearing in the text.

The paper uses the LDA algorithm to extract comment topics. The LDA topic model is an unsupervised machine learning technology based on the bag-of-words model [9], The biggest feature of the model is that the text-topic probability distribution and topic-word probability distribution obeys the Dirichlet distribution. However, due to the coarser granularity of the topic extraction of the LDA algorithm, the effect is poor in short texts such as movie reviews, so this paper proposes a topic extraction method combining improved k-means and TextRank.

Using the k-means algorithm to cluster the text vector matrix obtained after preprocessing. According to the distance between the data object and the cluster centroid, the object is assigned to the nearest cluster [10]. However, the k-means algorithm is sensitive to the initial clustering center, while the traditional k-means algorithm selects the initial clustering center randomly, so the clustering results often fluctuate. This paper adopts a density-based initial center selection method. By calculating the density of each region, the k points in the most distant high-density regions are selected as the initial clustering centers to improve the k-means algorithm. The density is expressed by the average distance of each point in the area, and the calculation formula is as follows.

$$\begin{cases} density = \dfrac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d(x_i, x_j) \\ d(x_i, x_j) = \| x_i - x_j \| \end{cases} \tag{1}$$

After using k-means to obtain different categories, using the TextRank algorithm to extract the overall keywords of each category as the topic of the category. The algorithm is based on the PageRank algorithm, constructs a network through the adjacent relationship between words, calculates node weight, and selects the T words with the highest weight as keywords through weight ranking [11]. The weight calculation formula is as follows.

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{2}$$

PCA algorithm uses the eigenvalues of the matrix to determine the dimension of the principal component and uses the eigenvector of the matrix to determine the direction of the principal component, finally converts the data set from the original space to the principal component space. Using this algorithm to reduce the vectorized data to easy to visualize it.

This paper analyses sentiment changes by calculating the sentiment score of the text. After segmenting the comment text, using Bayesian model training to predict the sentimental tendency [12,13] of the comment and mapping it to a 0-1 rating interval. Combining emotional scores with time and users' scores, comprehensively analysing the emotional changes in film reviews.

This paper according to the users' behavior preferences obtained after processing, combined with the user's rating of different movies to discuss the differential impact of the users' behavior preferences on the score. Since the scoring data is highly discretized and the changing trend is not obvious, linear regression is used to fit the data to highlight its changing laws.

3. Result

Using the methods discussed in chapter 2, this chapter analyses the review data, which is based on the obtained movie, and mainly discusses the results and significance of data mining.

The segmented words are divided into several groups of data according to the sentence, and the Apriori algorithm is used to analyse their relevance. The analysis results reflect the words which are often used in combination. The result is represented by an undirected network graph, as showed in Fig. 2. It can be seen from the results that the comments mainly focus on the three aspects of the expression of patriotic emotion, the review of personal experiences and national progress, and the film director.

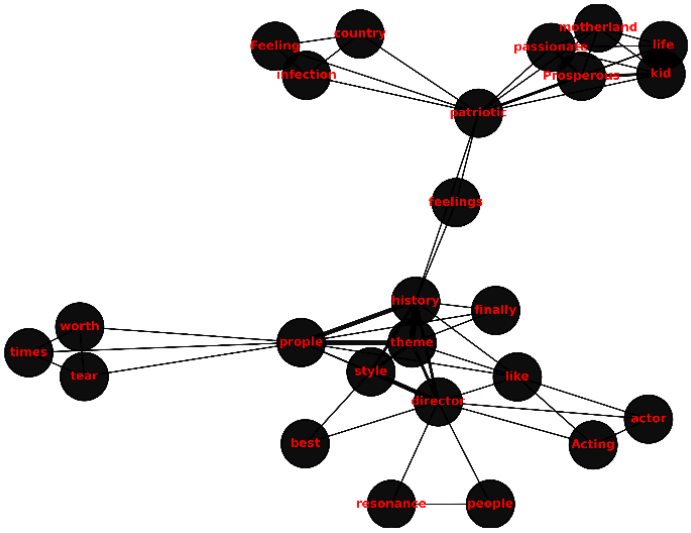


Fig.2 Undirected graph of related words

After the text is vectorized, the topic of the text is analysed by using the LDA algorithm, and the text is classified, the clustering result is shown in Fig.3. The results show that although the LDA algorithm can cluster texts according to topics, the topic extraction granularity is too coarse, resulting in a very large correlation between topics, and it cannot obtain meaningful information.

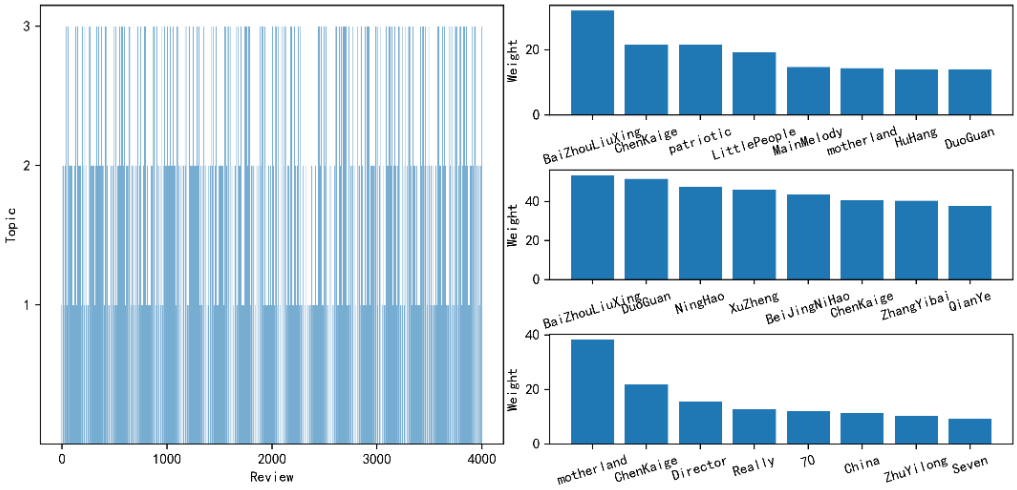


Fig.3 LDA clustering results

The topic is extracted using the improved k-means combined with the TextRank algorithm, and the visualized result after PCA dimensionality reduction is shown in Fig.4. The results show that the comment topic mainly focuses on the three aspects of film director, film story and the meaning of film expression. Among them, the average score of film director reviews is the lowest, and the score of film meaning is the highest, indicating that the audience more agrees with the connotation expressed by the film.

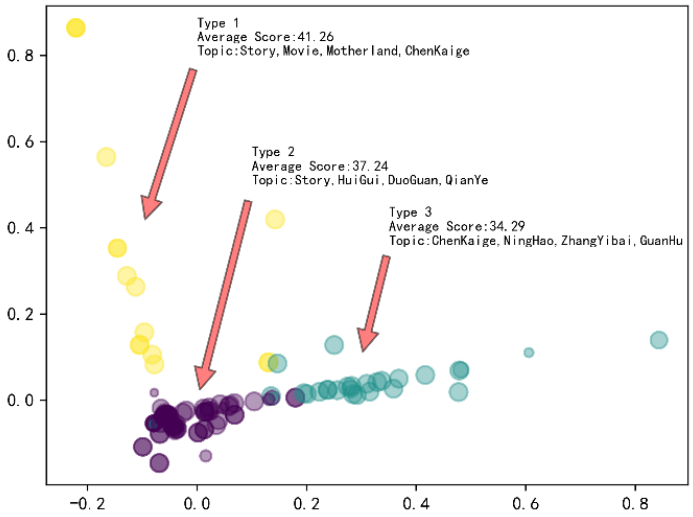


Fig.4 Topic extraction and clustering.

The sentiment analysis of the comment is performed to obtain the sentiment score of the comment, and the number of likes of the comment is calculated as the degree of support. These two indicators are combined with time and score to analyse the trend of public opinion. The result is expressed in a bubble chart, and the size of the bubble indicates the degree of support of the comment. The result is

shown in Fig.5.The results show that most of the comments are posted on the day the movie is released; the comments gradually decrease, after the movie is released one week; the support of high-scoring comments is far greater than the support of low-scoring comments; most of the comments are positive emotional tendencies.

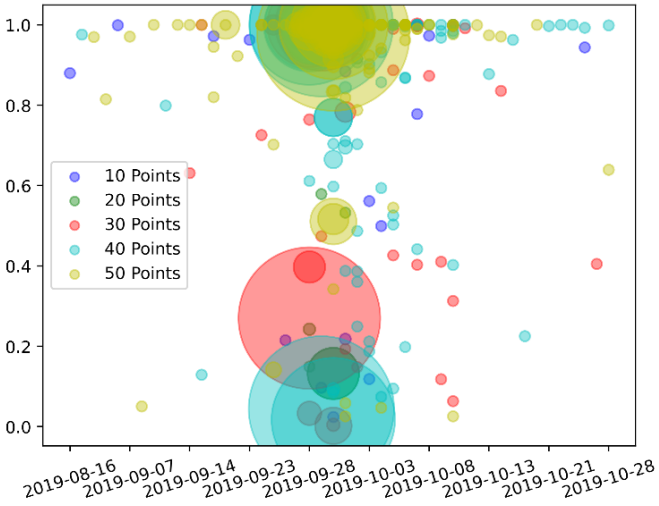


Fig.5 Sentiment analysis

Based on the user's previous movie watching data, the audience's movie-watching behavior preferences are summarized, and the audience's attention to domestic movies is quantified as a score between 0 and 1. Calculating the average score of different movies coming from the viewers with different behavior preferences, and analysing the influence of users with different behavior preferences on movie scores. Since the average score value is a discrete value and the transformation trend is not obvious, linear regression is used to fit the score data to make the law to be expressed more obvious. The result is shown in Fig.6 Here is a comparative analysis of the data of 5 movies, and the results clearly show the result: As users pay more attention to Chinese movies, the average score of Chinese movies by users increases accordingly, while the average score of other countries' movies decreases. This result fully shows that there is a huge correlation between the score of movies from users and the users' viewing habits.

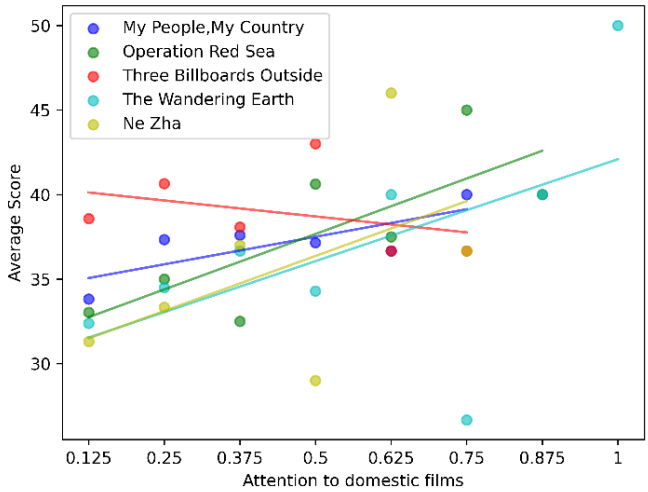


Fig.6 Behavioral preferences and ratings

4. Conclusion

This research uses movie reviews as a data set, extracts common word collocations in reviews based on the Apriori algorithm, extracts topics and clusters text based on the LDA model, and analyses the emotional tendency of reviews based on the Bayesian model. According to the users' viewing habits, analysing the influence of habits on the score of different movies from users. At the same time, due to the poor analysis effect of the LDA model on short texts, this paper proposes a topic model based on the improved k-means and the TextRank algorithm to effectively analyse the topic of the review. The final analysis result believes that the audience has disputes about the director and plot of the film "My People, My Country", but they generally agree with the connotation expressed by the film, and there are a lot of related topics about loving the motherland in the comments. In addition, the analysis also shows that the audience's movie-watching habits will affect their evaluation of movies from different sources, and audiences who prefer to watch Chinese movies will also rate Chinese movies higher.

This paper uses many models and methods to explore the semantic information of text and proposes its own topic model. It verifies the application of machine learning in data mining and provides a set of feasible processing templates for related research. In the future, the combination of various methods will be optimized, and more methods will be considered to improve the accuracy and effectiveness of the model.

References

[1] Zhuang L, Jing F and Xiaoyan Z 2006 Movie review mining and summarization *C. Proceedings of the 15th ACM international conference on Information and knowledge management* pp 43-50

[2] Yessenov K and Misailovic S 2009 Sentiment analysis of movie review comments *J. Methodology* **17** pp1-7

[3] Abulaish M, Jahiruddin, Doja M N and Ahmad T 2009 Feature and opinion mining for customer review summarization *C. International Conference on Pattern Recognition and Machine Intelligence* pp 219-214

[4] KTopal K and Ozsoyoglu G 2016 Movie review analysis: Emotion analysis of IMDb movie reviews *C. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* pp 1170-1176

[5] Thet T, Na J-C and Khoo CSG 2010 Aspect-based sentiment analysis of movie reviews on discussion boards *J. Journal of information science* **36(6)** pp 823-848

[6] Holt J D and Chung S M 1999 Efficient mining of association rules in text databases *C. Proceedings of the eighth international conference on Information and knowledge management* pp 234-242

[7] Yuan X 2017 An improved Apriori algorithm for mining association *C. AIP conference proceedings* **1820(1)** 080005

[8] Donghwa K, Deokseong S, Suhyoun C and Pilsung K 2019 Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec *J. Information Sciences* **477** pp 15-29

[9] Qiuxing C, Lixiu Y and Jie Y 2016 Short text classification based on LDA topic model *C. International Conference on Audio, Language and Image Processing (ICALIP)* pp 749-753

[10] Shyr-Shen Y, Shao-Wei C, Chuin-Mu W, Yung-Kuan C and Ting-Cheng C 2018 Two improved k-means algorithms *J. Applied Soft Computing* **68** pp 747-755

[11] Yujun W, Hui Y and Pengzhou Z 2016 Research on keyword extraction based on word2vec weighted textrank *C. 2016 2nd IEEE International Conference on Computer and Communications (ICCC)* pp 2109–2113

[12] Muhammad B, Huma I, Muhammad S and Amin K 2016 Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques *J. Journal of King Saud University-Computer and Information Sciences* **28(3)** pp 330-344

[13] Gutiérrez L, Bekios-Calfa J and Keith B 2018 A review on Bayesian networks for sentiment

analysis *C. International Conference on Software Process Improvement* pp 111-120