A Study of Sentiment Analysis of Moviegoing Reviews and Box Office Performance in the Chinese Comedy Film Market

Yitao Yin *

School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing, China, 210000

* Corresponding Author Email: 202213870009@nuist.edu.cn

Abstract. With the advancement of globalisation and digitalisation, the Chinese film market is rapidly becoming a force to be reckoned with in the global film industry. The purpose of this study is to investigate how emotional themes and the richness of emotional expression in film reviews afiect viewers' viewing decisions and thebox office performance of films. The average distribution and sentiment scores of each theme in the reviews of different movies are calculated by LDA theme model and sentiment analysis to measure the potential themes of emotional tendency, and the percentage of some specific emotion words in the reviews of different movies are calculated by 'WenXin' Chinese psychoanalysis system to measure the richness of emotional expression. Meanwhile, we define the box office retention rate.i.e.the box office of the second week divided by the boxoffice of the first week, to measure the influence of the potential themes of emotional tendency and the richness of emotional expression on the audience's decision to watch a flm. Through empirical research, we found that the richness of emotional expression in reviews of comedy films has a slight correlation with the audience's decision to watch the film, and the emotional richness of 'love' in reviews has a significant effect on the decision to watch the film.

Keywords: underlying theme, film review, sentiment analysis, box office performance, comedy film.

1. Introduction

1.1. Background to the study

In the wave of globalisation and digitalisation, the film market has become an important platform for cultural exchange and economic growth. Between 2010 and 2019, China's film market experienced remarkable growth, with box office revenue increasing more than sixfold from 10.172 billion yuan in 2010 to 64.266 billion yuan in 2019.At the onset of 2019, the emergence of the COVID-19 pandemic, often referred to as the New Crown Epidemic, precipitated a marked downturn in box office earnings and a substantial drop in cinema goers.

Fast forward to 2023, the Chinese film industry witnessed an annual box office revenue of 54.915 billion yuan, representing an impressive 83.4% increase compared to the previous year. This surge not only signifies a swift rebound of the Chinese film market but also underscores the substantial latent potential and expansive growth prospects that the market holds. Nevertheless, the intense competition in the market and the growing diversity of audience preferences present new challenges for film production and distribution. The influence of film reviews on audience decisions is escalating, and they play a crucial role in determining box office success.

1.2. Importance of the study

A thriving film market has far-reaching implications for the expansion of cultural industries, economic growth, job creation and the preservation of cultural diversity. Film reviews not only inform consumers, but also play a key role in shaping box office performance. A Consumer Reports survey found that less than 10 per cent of moviegoers do not consult film reviews and ratings at all before deciding to see a film. Audience reviews and feedback are directly correlated with market performance, especially in social media, where word-of-mouth can have a significant impact on the decision-making of potential viewers. In recent years, audience preferences in the Chinese film market have shown a trend of diversification and individualisation. Emotional resonance, the

experience of wonderful visual effects and the acting skills of the actors have become important criteria for them when choosing films. According to the "2024 Report on the Changing Trends of Chinese Moviegoers" published by the China Filmmakers Association in cooperation with the Lighthouse Research Institute, the most popular film genres among audiences include comedy, suspense, science fiction and action, with comedy genres standing out in terms of the number of releases and box office revenue.

1.3. Research questions

Movie reviews are an important reference point for viewers to make movie-watching decisions. A survey by Consumer Reports showed that less than 10% of viewers do not refer to movie reviews and ratings at all before choosing to watch a movie [1].

This study focuses on the impact of emotional disposition and richness of emotional expression in film reviews on viewers' decision making and film box office receipts. The study will examine the impact of the themes underlying emotional disposition and the richness of emotional expression on viewers' likelihood of watching a film.Using thematic modelling and sentiment analysis techniques, this study will quantify the textual data, identify the focus of public discussion and key features of the film, and reveal the general perceptions and concerns of the audience. This study is expected to provide a reference for the film industry and help develop more effective marketing strategies.

2. Literature review

2.1. Internet Word of Mouth

The rapid development of the Internet has given rise to the concept of Internet Word of Mouth (IWOM), which mainly refers to the personal emotions and passions that viewers share through online platforms after watching a film, and these shares have a significant impact on potential viewers' decision to watch a film [2][3].

In 2006, Liu Yong highlighted that the initial word of mouth for a film within its first few weeks of release can significantly influence the box office performance. Furthermore, he noted that the weekly box office results of a film are closely mirrored by the film's word of mouth dynamics (Liu, 2006). Jordi McKenzie pointed out in her 2009 study that box office and word-of-mouth of a movie positively influence each other [5]. With the widespread use of mobile internet, the dissemination of IWOM is mainly realised in the form of online reviews [6]. With the proliferation of new media, the influence of IWOM fored by viewers through online media such as professional forums and film review websites on the public's decision to watch films can no longer be ignoredand in some cases has even surpassed the publicity effect of traditional mainstream media[7]. Zhu Rui et al. [10][11] used web crawler technology to collect review data of 160 films, quantified the review content through sentiment analysis, and applied time series model analysis, and found that the sentiment trend in reviews had a positive impact on the box office of films [10][11].

The origins of Internet Word of Mouth (IWOM) have become increasingly varied, with online textual reviews emerging as a pivotal element in the IWOM's development. This is particularly true as consumers are more inclined to consider negative feedback when making decisions about film viewing, highlighting the significant role that IWOM plays in shaping public opinion and influencing box office performance.

The film industry is gradually moving into a new era that is oriented towards word-of-mouth [11]. Therefore, whether from the perspective of shaping the word of mouth of a film or studying the decision-making tendency of the audience, movie reviews, as an important form of expression of online word of mouth, are of great research value and practical significance.

2.2. Persuasion effect

Ma Xiangyang et al. pointed out in 2012 that the persuasion effect refers to the phenomenon that when faced with persuasive information, an individual's attitude changes and influences his decision-making behaviour which is common in life consumption and has important application value. The emotion contained in the information source is one of the main factors affecting the persuasion effect, and the emotion has a non-negligible influence on people's behaviour evaluation and decision-making[12]. In 2018, Chi Jianyu and Luo Ziheng empirically analysed the relationship between word-of-mouth and film box office using a single equation model and a joint equation model, and found that IWOM has a significant positive influence on film box office in China. This impact works through the knowledge effect and persuasion effect, especially in the first two weeks when the impact is most significant [13].

Consequently, interactive word of mouth (IWOM) exerts a substantial positive influence on film box office sales, primarily through the emotional transmission and the persuasive power of the information it conveys.This finding highlights the importance of considering the audience's emotional expression in film marketing, and also provides empirical support for decision making in the film industry.

3. Methodology

3.1. LDA Modeling

3.1.1. Text vectorization

To enable computers to process text data, it is essential to convert text information into numerical vectors that capture the semantics of the text. This conversion allows text data to be transformed into a format recognizable by computers. In this study, we employ the bag-of-words model and the TF-IDF method to vectorize the text's feature words.

3.1.2. Bag-of-words model

The bag-of-words model serves as a straightforward approach to text representation, facilitating the transformation of textual content into a numerical vector based on word frequencies. This model enables machine learning algorithms to analyze text data by quantifying the frequency of each word's appearance. It disregards the sequential arrangement of words, viewing each document as an unordered set of words and tabulating the frequency of each term's occurrence.

The basic idea of bag-of-words modeling is to segment the text into individual words and then place these words into a large word list. For each text, a corresponding word frequency vector is generated, and each element within this vector corresponds to the frequency with which a particular word is encountered throughout the text. The model mathematical formula can be expressed as:

$$\vec{\mathrm{d}}_{\mathrm{j}} = (\mathrm{n}_{1\mathrm{j}}, \mathrm{n}_{2\mathrm{j}}, \cdots, \mathrm{n}_{\mathrm{ij}}, \cdots \mathrm{n}_{\mathrm{mj}})^{\mathrm{T}}, \mathrm{j} = 1,2, \cdots, \mathrm{D} \quad (1)$$

Where $\vec{d}_j$ denotes the vector associated with document$j$;m signifies the total number of unique terms in the dictionary;D indicates the total number of documents within the sample; and n_{ij} is the frequency of word i occurring in document j.

Typically, the vocabulary utilized within a given document is considerably more limited compared to the extensive range of words found in a comprehensive dictionary. Consequently, the bag-of-words model is employed to encapsulate the representation of documents, which are inherently characterized by a high degree of sparsity, as illustrated in the subsequent example:

$$\mathrm{CV}_{\mathrm{D}\times\mathrm{m}} = (\vec{\mathrm{d}}_1 \quad \cdots \quad \vec{\mathrm{d}}_{\mathrm{D}})^{\mathrm{T}} = \begin{pmatrix} \mathrm{n}_{11} & \cdots & \mathrm{n}_{\mathrm{m}1} \\ \vdots & \ddots & \vdots \\ \mathrm{n}_{1\mathrm{D}} & \cdots & \mathrm{n}_{\mathrm{mD}} \end{pmatrix} \quad (2)$$

3.1.3. TF-IDF

Widely recognized in the realms of information retrieval and text mining, the Word Frequency-Inverse Document Frequency (TF-IDF) technique is a fundamental approach to assigning weights to terms. This method is frequently employed to gauge the prominence of a term within a document in contrast to its prevalence across the broader collection of documents. The TF-IDF framework quantifies the importance of each term within a document by calculating its TF-IDF score against a comprehensive lexicon, which is instrumental in determining the term's relevance across multiple documents. Comprising two essential elements, the TF-IDF algorithm hinges on the concepts of term frequency and inverse document frequency.

Word frequency is the frequency of occurrence of a word in a given text, all words are considered equally important and the formula is expressed as follows:

$$\mathrm{T_{i,j}} = \frac{\mathrm{n_{i,j}}}{\sum_{\mathrm{k}} \mathrm{n_{k,j}}} \tag{3}$$

Where$T_{i,j}$denotes the word frequency of the word$i$within document d_j,n signifies the total count of words,, $n_{i,j}$ represents the number of times the word i appears in the document d_j ,and the denominator indicates the overall word count present in document d_j .

Inverse document frequency is a measure of the general importance of a word in a document, and its size is inversely proportional to how common a word is. The fewer times the term appears in the document, the more representative and discriminating it is. We need to reduce the weight of frequent terms while expanding the weight of rare terms. The formula is expressed as follows:

$$\mathrm{I_i} = \log \frac{|\ \mathrm{D}\ |}{|\{\mathrm{j} = \mathrm{n_i} \in \mathrm{d_j}\}| + 1} \tag{4}$$

WhereI_i represents the frequency of reverse documents for the term$i$,$|D|$ is the number of documents in the corpus, and$|\{j = n_i \in d_j\}|$ is the number of documents containing the term. When the word does not belong to the corpus, the denominator will be 0. Mathematically, it is not possible to perform logarithmic operations, so the denominator is often expressed as$\left|\{j = n_i \in d_j\}\right| + 1$.

Finally, the TF-IDF value algorithm for a word is represented as:

$$\mathrm{TF-IDF} = \mathrm{T_{i,j}} \cdot \mathrm{I_i} \tag{5}$$

3.1.4. Determining the optimal number of topics

The identification of the most suitable number of topics within an LDA topic model is facilitated by evaluating the perplexity and coherence scores, which serve as key metrics for model performance.

Perplexity is an important indicator of how good a language model is, and it reflects the model's ability to predict the test data. The formula is as follows:

$$\mathrm{Perplexity}\ = \exp\left(-\frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \log \mathrm{P}(\mathrm{w_i}|\mathrm{w_1}, \mathrm{w_2}, \ldots, \mathrm{w_{i-1}})\right) \tag{6}$$

Where$N$ is the total number of words in the corpus,w_i is the$i$th word, and$P(w_i|w_1, w_2, \ldots, w_{i-1})$ is the probability that the model predicts the$i$ th word, given that the antecedent is known. The lower the perplexity, the better the prediction ability of the model.

The consistency score is used to measure the quality of topic models, especially in LDA models. It measures the semantic consistency of the words in the topic. The formula for calculating it is given below:

$$\mathrm{C_{UMass}} = \frac{2}{\mathrm{N} \cdot (\mathrm{N}-1)} \sum_{\mathrm{i=2}}^{\mathrm{N}} \sum_{\mathrm{j=1}}^{\mathrm{i-1}} \log \frac{\mathrm{P}(\mathrm{w_i}, \mathrm{w_j})}{\mathrm{P}(\mathrm{w_j})} \tag{7}$$

Where$N$ is the number of words in the theme,$P(w_i, w_j)$ is the probability of two words appearing at the same time, and$P(w_j)$ is the probability of the wordw_j appearing. An elevated consistency score is indicative of superior semantic coherence within the theme.

3.1.5. Model Construction

LDA topic model is a probabilistic model for text categorization and topic generation, which is an unsupervised machine learning technique that contains a three-layer structure of words, topics, and documents.The LDA model treats a document as a mixture of multiple topics, each of which represents a potential concept or topic in the document, and by analyzing the extracted topics in the document, it is possible to categorize the document into the corresponding topic categories, thus realizing topic clustering or text categorization. thereby realizing topic clustering or text classification. It can be visualized by the following probability formula:

$$\mathrm{P(words|documents)} = \mathrm{P(words|topics)} \times \mathrm{P(topics|documents)} \tag{8}$$

3.2. Sentiment analysis

In this paper, we use SnowNLP in python to calculate the sentiment score for each topic.The sentiment analysis model of SnowNLP predicts the sentiment tendency of a text by calculating the conditional probability of each word in the text under different sentiment categories and combining it with the prior probability and feature independence assumptions. In SnowNLP, the probability calculation for sentiment analysis can be expressed as:

$$\mathrm{P(S_i|W)} = \frac{\prod_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{P}\left(\mathrm{W_j|S_i}\right)^{\mathrm{n_{W_j}}}}{\mathrm{P(W)}} \tag{9}$$

Where S_i is the sentiment category$i$ (positive or negative),W is the aggregate of words in the text,$\mathrm{W_j}$ is the$j$ th word in the text,n_{W_j} is the number of times the wordW_j appears in the text, and$P(W_j|S_i)$ is the conditional probability of the word W_j given the sentiment S_i .

3.3. Sentiment Segmentation

In this study, we utilized the "Wenxin" Chinese psychoanalysis system to segment the comments and select emotion-related features, including Affect, PosEmo, NegEmo, Anx, Anger, Sad, Swear, Achieve, Leisure, Love, and NumEmotion.

4. Result

4.1. Data sources

In this paper, we use the kaggle dataset "CMM (Chinese Multi-modal Movie)", which contains 921 movies released between January 2017 and March 2024 and more than 500,000 user reviews. We first screened 193 comedy movies released during non-holiday periods from January 2017 to March 2024 and 47,898 reviews from the first week of their release.

4.2. Determination of the optimal number of themes

The optimal number of topics for constructing the LDA topic model was determined to be 10 based on the perplexity and consistency scores (Figure 1).

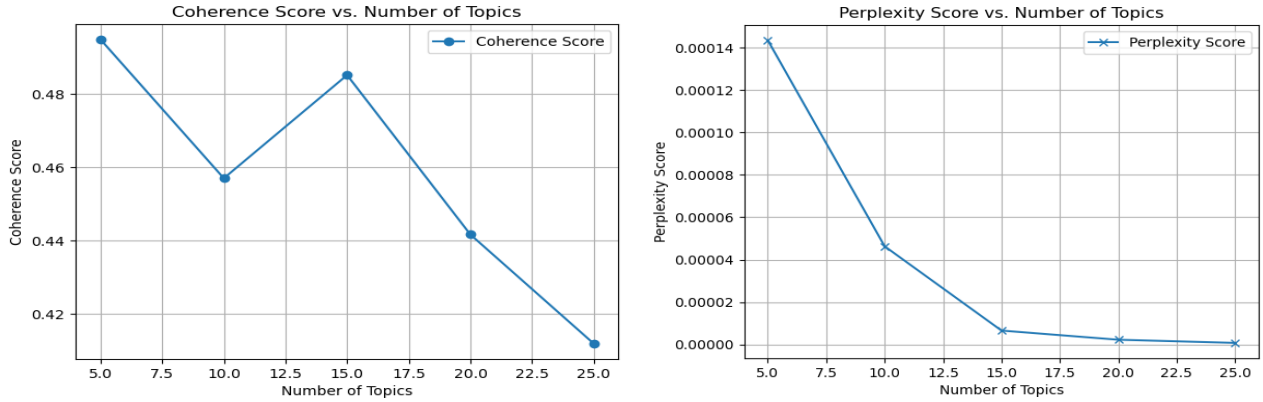

Figure 1. This caption has one line so it is centered.

4.3. The average distribution of each topic for each movie

Utilizing LDA topic modeling, this research performed a comprehensive topic analysis on 47,898 reviews of comedy films. The analysis successfully identified 10 distinct topics, determined the document-topic distribution for each topic, and calculated the average distribution of each topic across all movies (Table 1).

Table 1. The average distribution of each topic for each movie.

ID	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9
4840388	0.0344	0.0344	0.0368	0.0328	0.6742	0.0355	0.0364	0.0370	0.0449	0.0335
22232939	0.0395	0.0475	0.0492	0.0447	0.5783	0.0429	0.0488	0.0438	0.0661	0.0393
22266126	0.0467	0.0683	0.0514	0.0646	0.5463	0.0385	0.0401	0.0397	0.0589	0.0456
22557335	0.0507	0.0594	0.0608	0.0501	0.4861	0.0386	0.0834	0.0516	0.0723	0.0469
23774869	0.0466	0.0432	0.0579	0.0351	0.5774	0.0487	0.0419	0.0474	0.0582	0.0436

4.4. Topic Keywords and Sentiment Score

Based on the 10 potential themes obtained from the LDA theme model and the distribution of feature words for each theme, the top 8 high-probability lexical items are collated and ranked in descending order according to their probability of occurrence within the theme. Then, based on the high probability feature words under each theme, the most representative theme identifiers are summarised and generalised. Meanwhile, nowNLP in python is used to calculate the sentiment score of each topic, where the closer to 1 the more positive and the closer to 0 the more negative. According to the results, it can be seen that topic5 and topic1 are the topics with the lowest sentiment score (0.4797) and the highest sentiment score (0.7820), respectively, which indicates that the sentiment of topic5 is the most negative, while the sentiment of topic1 is the most positive (Table 2).

Table 2. Topic Keywords and Sentiment Score.

Topic	Keywords								Emotion Score
Topic0	Playing Around	Love to Watch	Awakening	Don't Want	Woody	Maodou (Bean)	Fox	Good Looking	0.5320
Topic1	Acting Skills	Feeling	Awkward	Good Looking	Not Bad	Like	Plot	Movie	0.7820
Topic2	Worse	Cemetery	Firm	Don't Watch	A While	Disgusting	Movie	Deceiving	0.4969
Topic3	Boycott	Honey Sweet	Two Stars	Song Qian	Hahaha	Crap	Release	Bad Movie	0.5269
Topic4	Zoo	Light Comedy	Disgusting	Otaku	Sister	Full Marks	Regret	One Point	0.5751
Topic5	Chasing the Dragon	To Take Over	Don't Make Trouble	Beautiful	Dream Chaser	Showbiz	So-So	Trash	0.4797
Topic6	Option	Peanut	Tomohisa Yamashita	Under the Mountain	Godlike	none	Thing	Negative Score	0.5040
Topic7	Mom	13	Downey	Not Bad	TuTu	Hmmm	Hahaha	Cute	0.6651
Topic8	Cake Face	Where	Jaycee Chan	Shu Ke	Beta	Refund	Damn	Request for Refund	0.5495
Topic9	Zhang Yunlong	Share	Star	Wong Cho-lam	Really Bad	Dilraba	Thing	Reeba	0.5419

4.5. Regression analysis

Zhang Chi [14], in the article "A Study on the Impact of First-Week Word of Mouth on the Box Office of Domestic Films", explains that among the factors influencing the box office of a film, firstly, the first-week word of mouth of a film is the most important factor influencing the total box office of a film; Secondly, the variable of word-of-mouth during the initial week predominantly influences the box office revenues in the period following the first week. It is observed that negative sentiment exerts a more substantial effect on box office outcomes than positive sentiment. This observation suggests that individuals may be more inclined to give credence to negative critiques when making decisions regarding their cinema visits [14]. This paper defines the film's box office maintenance rate, i.e.. the second week's box office divided by the first week's box office, as the dependent variable of the regression analysis model..Emotion-related features (emotion, positive emotion, negative emotion, anxiety, anger, sadness, oath, achievement, leisure, love, and number of emotions) and the average distribution of 10 themes, number of reviews and average of review words for each film obtained from the Chinese ‘Wenxin’ psychoanalysis system are used as independent variables.

The outcomes of the regression analysis reveal that the F-statistic stands at 4.721, with an associated P-value significantly below the 0.05 threshold, thereby confirming the overall statistical significance of the model. The model's fit is quantified by a coefficient of 0.379. In the assessment of individual variables, both the average word count and the 'love'-related emotional terms have P-values that fall below 0.05, successfully meeting the criteria for statistical significance. Conversely, the remaining independent variables exhibit P-values that exceed 0.05, signifying that their influence on the dependent variable y lacks statistical relevance.

5. Conclusion

Strongly emotional comments in comedy movie reviews can influence the audience's decision to watch the movie, in which the richness of the emotional expression of "love" has a significant impact on the audience's decision to watch the movie, which in turn affects the box office. The more information contained in the movie reviews, the more likely it is to influence the audience's decision to watch the movie, increase the box office retention rate, and then increase the box office

revenue..Future research can further refine the classification of emotions, consider various types of movies in a comprehensive and holistic way, and study the impact of different emotions on the box office of different types of movies.

References

[1] Chen, G. & Dai, T. T.. (2017). Analysis of Influencing Factors of Movie Consumption Decision. Modern Business (30), 20-21. doi:10.14097/j.cnki.5392/2017.30.007.

[2] Gasimli V, Jiang M, Yuan X, et al. The informational role of average rating and variance of customer ratings in the differential patterns of consumer behavior [J]. Human Systems Management, 2020, 39 (1): 1-10.

[3] Godes D, Mayzlin D. Using online conversations to study word-of-mouth communication [J].Marketing science, 2004, 23 (4): 545-560.

[4] Liu Y. Word of mouth for movies: Its dynamics and impact on box office revenue [J]. Journal of marketing, 2006, 70 (3): 74-89.

[5] Jordi McKenzie.(2009).Revealed word-of-mouth demand and adaptive supply: survival of motion pictures at the Australian box office.Journal of Journal of Cultural Economics (4),279-299.

[6] Dellarocas C, Zhang X, Awad N F. Exploring the value of online product reviews in forecasting sales: The case of motion pictures [J]. Journal of Interactive marketing, 2007, 21 (4): 23-45.

[7] SHI WEN-HUA,ZHONG BI-YUAN,ZHANG QI. A comparative study of the impact of online movie reviews and online short reviews on box office revenue [J]. China Management Science Science,2017,25 (10):162-170.DOI:10.16381/j.cnki.issn1003-207x.2017.10.017.

[8] Dan Lu,Yutian Fan. Research on the Influence of Weibo Marketing on Box Office of Films A Case Study of the Top 20 Films in the Chinese Box Office Rankings in 2019 [C]//.Proceedings of 6th International Conference on Economics, Management, Law and Education (EMLE)2020). ,2020:227-234.DOI:10.26914/c.cnkihy.2020.051742.

[9] Kejin Xu. Analysis of the impact of new media environment on domestic movie box office [J]. Western Radio and Television,2021,42 (17):155-157.

[10] ZHU Rui,MA Yongmei,CHEN Yanan. Research on Internet Word of Mouth in Movie Box Office Forecasting--An Empirical Analysis Based on Chinese Mainland Film Market [J]. Journal of Chaohu College,2021,23 (05):71-79

[11] Hu S, Li B, Chai J, et al. Study on Box Office Influence Factors During the Film's Release Period [C]//2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS). ieee, 2019: 338-342.

[12] Ma, Xiangyang, Xu, Fuming, Wu, Xiuliang, Pan, Jing & Li, Tian. (2012). Theoretical model, influencing factors and coping strategies of the persuasion effect. Advances in Psychological Science (05), 735-744.

[13] Ji, Jianyu & Luo, Ziheng. (2018). Persuasion effect and knowledge effect - An empirical study on the influence of IWOM on Chinese movie box office. China News Communication Research (01), 156-171.

[14] Zhang, Chi. (2020). A study on the impact of first-week word-of-mouth on box office of domestic movies (Master's thesis,Cheng du University of Technology). Master https://link.cnki.net/doi/10.26986/d.cnki.gcdlc.2020.001240doi:10.26986/d.cnki.gcdlc.2020.001240.