Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 7756896, 11 pages
https://doi.org/10.1155/2022/7756896

Research Article

Application of Cluster Analysis Technology in Visualization Research of Movie Review Data

Bin Xu, Cheng Chen, and Jong-Hoon Yang

Department of Digital Image in Sangmyung University, Seoul 03015, Republic of Korea

Correspondence should be addressed to Cheng Chen; 2016120808@jou.edu.cn

Received 14 May 2022; Accepted 15 June 2022; Published 16 July 2022

Academic Editor: Qiangyi Li

In order to improve the reference value of film review data, this paper combines clustering analysis technology to construct a film review clustering visualization research system to improve the visualization effect of film reviews. Moreover, this paper analyzes online reviews based on the method of type-2 fuzzy sets and determines the interval of type-2 fuzzy sets of each factor set of commodity status in commercial stores by considering both the language and specific numbers of keywords in online review texts. In addition, this paper analyzes intelligent control through practical application and sets the expected target of the control object. Finally, according to the fuzzy rules and word calculation, this paper determines the interference or improvement measures that can be referenced to form a closed-loop control. The experimental research shows that cluster analysis technology can play a certain role in the visual analysis of film review data.

1. Introduction

With the rapid development of the Internet, film review data has become more and more open and diverse, and it has attracted more and more scientists' attention. In particular, on film websites, unconstrained users freely express their personal opinions and comments, which also forms a certain cultural phenomenon [1]. Moreover, film review data tend to be more appreciative and promotional, and film review data also tend to be different and fluid. Due to the popularity and prosperity of online media, the analysis of film reviews has been pushed to a climax [2], and it is far from enough to analyze film review data through human perception of film reviews. Therefore, more and more leaders in the computer industry conduct data analysis by processing and converting data, and making the data clearer and clearer through visualization, and the analysis results are deeply rooted in the hearts of the people [3].

By using visualization technology to visualize film review data, researchers break the "black box" of traditional analysis data [4], which can greatly enhance users' trust in analysis results. Visualization technology application methods mainly include the following aspects. First, in the data preprocessing stage, the data are directly reflected in the form of graphics to provide users with an intuitive impression, which can not only help users to understand the data from a macro perspective [5] but also help users determine the direction of analyzing the data. Second, in the data analysis stage, after the film review data are subjected to word segmentation, word frequency extraction, and other operations, these words can be displayed in a visual way. Moreover, it intuitively allows users to judge the graphics, deepens the user's understanding of the data, and enhances the interaction between data analysis and human knowledge, so it is a perfect fusion of data analysis and visualization technology. Third, in the result display stage, the analysis results of film review data are used to generate understandable graphs using visualization techniques [6]. In addition, the analysis process can also be displayed in a graphic image, which allows users to have a deep and intuitive understanding, and also makes the data analysis results clearer.

A good movie usually has many elements, such as a reasonable plot, superb acting, grand scenes, and shocking sound effects. These influencing factors are also frequently discussed in film reviews (especially long-form reviews). In

order to enable our system to better understand the content of long film reviews, and more accurately identify which aspect of the film a certain long film review text discusses. We need to establish a relatively complete set of movie review keywords, such as the three keywords of plot, story, and events should correspond to the element of "plot" [7]; and the three keywords of sound effects, sound, and music should correspond to "sound" element. Although the work of establishing a keyword list for film reviews can be achieved manually, because Chinese has a large number of vocabulary, it is tantamount to looking for a needle in a haystack. In addition to the richness and complexity of Chinese expressions, it is difficult for us to manually find all the corresponding elements in keywords. Therefore, we need to use computer technology to help retrieve massive vocabulary and find these target keywords [8].

The class is to divide a data set into different classes or clusters according to a certain standard (such as a distance criterion, that is, the distance between data points) so that the similarity of data objects in the same cluster is as large as possible, and at the same time they are not in the same cluster. The data objects in a cluster are also as diverse as possible. We can specifically understand that after clustering, the data of the same class should be clustered together as much as possible and the data of different classes should be separated as much as possible [9]. The research of cluster analysis mainly focuses on distance-based clustering, fuzzy relation-based clustering, and objective function-based clustering. In the field of data mining, clustering methods include statistical methods, machine learning methods, and database-oriented methods [10].

The keywords of text are the smallest units used to understand the text. By extracting keywords that are closely related to the theme and connotation of the text, the meaning of the text expression can be grasped more quickly. In this paper, a machine learning algorithm is used for text processing of large-scale movie review text data. First, keyword extraction technology is performed, and then the movie review data are visually analyzed, so as to deeply and clearly describe the theme of the review and the user's emotional tendency [11]. The beginning of the research and application of text keyword extraction technology is marked by the automatic document tagging method based on word frequency statistics proposed by IBM Corporation of the United States. Since then, it has attracted the attention and exploration of many domestic and foreign researchers and formed several major categories, such as machine learning analysis methods, statistical analysis methods, and network analysis methods [3].

The keyword extraction process can be regarded as a binary classification problem to determine whether a word is a keyword, and the requirement is to establish a classification model based on features. Commonly used models are decision tree (DT), Naive Bayes (NB), support vector machine (SVM), hidden Markov model (HMM), conditional random field (CRF) model, etc. Turneyt proposed the Genex model, which uses a decision tree algorithm as a classifier and uses word frequency and parts of speech as features. Frankt proposed the KEA model, which implements the function of automatically extracting keywords using a Naive Bayes classifier [12]. These two models together become the benchmark system for supervised method models for automatic keyword extraction. The algorithm model of supervised learning often needs the support of a stable large-scale corpus during training, which is a large amount of data materials stored in the computer that are real and effective and have specific annotations [13] using statistical analysis methods to extract keywords. It is simpler to say and easier to understand. Theoretically, no training data and knowledge base are needed, and a collection of keywords can be obtained by processing information such as word frequency statistics, part-of-speech filtering, mean, and variance, through artificially set rules. There are three commonly used statistical and quantitative indicators for keyword judgment, namely, word-vector-based statistical and quantitative indicators; word-based document location statistical and quantitative indicators; and word-based associated information statistical and quantitative indicators.

Online reviews are a form of word-of-mouth. The storability and ease of processing of word of mouth information make word-of-mouth spread longer. Due to the increasing influence of word-of-mouth on consumers in the online environment, many scholars have begun to explore the impact of word-of-mouth on performance at the macro-group level. [14]. Relevant research can be summarized as three levels of research: the impact of word-of-mouth on corporate operating income, the impact of word-of-mouth on the promotion of new movies, and the impact of word-of-mouth on corporate stock value. However, the specifics that then affect the company's revenue, the promotion of the movie, and the value of the stock are all based on the analysis of word-of-mouth, the language of paper reviews. The influencing factors of the content of online comment information include the comprehensive evaluation of the event itself and are also affected by factors such as self-orientation, community orientation, business orientation, and reciprocity motivation [15]. It has been mentioned in the research on the effectiveness of online reviews that its effectiveness includes four dimensions, information content, information sources, movie types, and publication time. Therefore, in order to analyze the content of online reviews, it is necessary to discuss many factors and dynamic dimensions of online review information content [16].

This paper combines the cluster analysis technology to construct a film review cluster visualization research system, improve the film review visualization effect, promote the film to be better understood by people, and make it easier for people to choose excellent films to watch.

2. Comprehensive Evaluation and Language Dynamics Analysis of Film Reviews under IT2FS

2.1. Analysis of Film Review Language Based on Type 2 Fuzzy Sets. This paper discusses an analysis of online reviews based on the same type of film. According to the practice process, and survey statistics and cluster analysis are calculated. We

assume that $\Omega_n = \{u_{n1}, u_{n2}, \ldots, u_{nm}\}$ is a discrete time-varying universe, where u_{ni} (i = 1, 2, ...) represents the same film in the ith theater in the nth film evaluation. On the domain of discourse Ω_n, we assume that the five base words NB, NS, ZO, PS, and PB represent the nth satisfaction evaluation "very dissatisfied," "dissatisfied," "average," "satisfied," and "very satisfied," respectively, the corresponding fuzzy sets are ω_1^n‘ω_2^n‘ω_3^n‘ω_4^n, and ω_5^n, respectively. $\omega_k^n(u_{ni})$ represents the degree to which the scores of similar films in the ith theater in the nth satisfaction evaluation belong to the fuzzy set $\omega_k^n(k = 1, 2, 3, 4, 5)$, as shown in Figure 1.

After applying the division, its membership function can be expressed as

$$\omega_{jk}^n = \int_{u\in\Omega}\int_{x\in L_{ujk}^n} \frac{\mu_{\omega_{jk}^n}(u,x)}{(u,x)}, L_{ujk}^n \subseteq [0,1] \text{and} \, \mu_{\omega_{jk}^n}(u,x) = 1. \tag{1}$$

Among them, there are

$$\begin{aligned} L_{ujk}^n &= \left[\underline{g_{jk}^n}, \overline{g_{jk}^n}\right], \\ \text{FOU}\left(\omega_{jk}^n\right) &= \underset{u\subset\Omega}{\cup} u \times L_{ujk}^n. \end{aligned} \tag{2}$$

Among them, $L_{ujk}^n = [g_{jk}^n \overline{g_{jk}^n}]$, the lower boundary $g_{jk}^n \overline{g_{jk}^n}$ is based on FOUK as shown in the following equation, and its functions can be expressed as follows:

$$\underline{g_{j1}^n} = \begin{cases} 1 & 0 \le z_{ji}^n < 1 \\ \dfrac{1.5 - z_{ji}^n}{0.5} & 1 \le z_{ji}^n < 1.5, \\ 0 & 1.5 \le z_{ji}^n \le 5 \end{cases} \underline{g_{j2}^n} = \begin{cases} 0, & 0 \le z_{ji}^n \le 1.5 \\ \dfrac{z_{ji}^n - 1.5}{0.5}, & 1.5 < z_{ji}^n \le 2 \\ \dfrac{2.5 - z_{ji}^n}{0.5}, & 2 < z_{ji}^n \le 2.5 \\ 0, & 2.5 < z_{ji}^n \le 5 \end{cases},$$

$$\underline{g_{j3}^n} = \begin{cases} 0, & 0 \le z_{ji}^n \le 2.5 \\ \dfrac{z_{ji}^n - 2.5}{0.5}, & 2.5 < z_{ji}^n \le 3 \\ \dfrac{3.5 - z_{ji}^n}{0.5}, & 3 < z_{ji}^n \le 3.5 \\ 0, & 3.5 < z_{ji}^n \le 5 \end{cases} \underline{g_{j4}^n} = \begin{cases} 0, & 0 \le z_{ji}^n \le 3.5 \\ \dfrac{z_{ji}^n - 3.5}{0.5}, & 3.5 < z_{ji}^n \le 4 \\ \dfrac{4.5 - z_{ji}^n}{0.5}, & 4 < z_{ji}^n \le 4.5 \\ 0, & 4.5 < z_{ji}^n \le 5 \end{cases},$$

$$\underline{g_{j5}^n} = \begin{cases} 0, & 0 \le z_{ji}^n \le 4.5 \\ \dfrac{z_{ji}^n - 4.5}{0.5}, & 4.5 < z_{ji}^n \le 5^{\circ} \end{cases} \tag{3}$$

According to the FOU of the following, the upper boundary $\overline{g_{jk}^n}$ can be expressed as

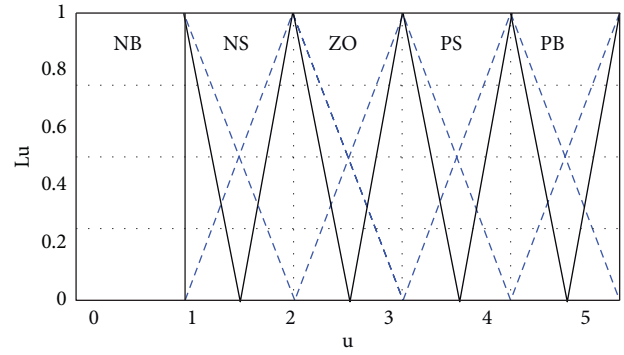

Figure 1: FOU of the base word membership function.

$$\overline{g_{j1}^n} = \begin{cases} 1 & 0 \le z_{ji}^n < 1 \\ 2 - z_{ji}^n & 1 \le z_{ji}^n < 2 \\ 0 & 2 \le z_{ji}^n \le 5 \end{cases}, \overline{g_{j2}^n} = \begin{cases} 0, & 0 \le z_{ji}^n \le 1 \\ z_{ji}^n - 1, & 1 < z_{ji}^n \le 2 \\ 3 - z_{ji}^n, & 2 < z_{ji}^n \le 3 \\ 0, & 3 < z_{ji}^n \le 5 \end{cases},$$

$$\overline{g_{j3}^n} = \begin{cases} 0, & 0 \le z_{ji}^n \le 2 \\ z_{ji}^n - 2, & 2 < z_{ji}^n \le 3 \\ 4 - z_{ji}^n, & 3 < z_{ji}^n \le 4 \\ 0, & 3.5 < z_{ji}^n \le 5 \end{cases}, \overline{g_{j4}^n} = \begin{cases} 0, & 0 \le z_{ji}^n \le 3 \\ z_{ji}^n - 3, & 3 < z_{ji}^n \le 4 \\ 5 - z_{ji}^n, & 4 < z_{ji}^n \le 5 \end{cases},$$

$$\overline{g_{j5}^n} = \begin{cases} 0, & 0 \le z_{ji}^n \le 4 \\ z_{ji}^n - 4, & 4 < z_{ji}^n \le 5 \end{cases}, \tag{4}$$

where z_{ji}^n is determined by the total number of online comments, the number of negative comments, the number of moderate comments, the number of general comments, the number of positive comments, and the number of very positive comments. Its calculation method is as follows.

2.1.1. Text Keywords Are Language Words. The domain of discourse performs part-of-speech tagging and word frequency statistics on the comment language in the t_n time period, and extracts text keywords based on the results of word frequency statistics. We assume that there are a total of N_{t_n} comments in the t_n time period, a total of $N_{t_n 1}$ comments in the middle, and a total of $N_{t_t 2}$ comments in the negative comments. Keywords are counted. Through the cluster analysis method in Section 3, it can be obtained that the descriptions of each influencing factor are $a_{t_n/1}(j = 1, 2, 3, \ldots)$ items for "bad review," $a_{t_k/2}$ items for "moderate review," $a_{t,j4}$ items for "good review," and $a_{t_n j5}$ items for "very good review." According to the actual data statistics and analysis results, the semantic scale of the comment sets are named "bad comment," "moderate comment," "general comment," "good comment," and "excellent," respectively." Assigned as $p'_{t_k/1}, p'_{t_n/2}, p'_{t_n\beta 3}, , p'_{t_k,4}$, $p'_{t_n j5}$ accordingly.

The composition weight vector is as follows:

$$P_{t_nj}{}' = \left(p_{t_nj1}{}', p_{t_n/2}{}', p_{t_n\beta 3}{}', p_{t_{nj}}{}', p_{t_nj5}{}' \right). \tag{5}$$

Using the weighted average fuzzy set operator to synthesize the weight vector $P_{t_nj}{}'$ and the online review evaluation matrix $R_{t_n,i}{}'$, we can get

$$Z_i^n = P_{t_nj}{}' \bullet R_{t_ni}{}' = \left(z_{1i}^n, z_{2i}^n, z_{3i}^n, \ldots, z_{ji}^n, \ldots \right). \tag{6}$$

Among them, there are

$$R_{t_ni}{}' = \begin{pmatrix} a_{t_n11} & a_{t_n21} & \cdots & a_{t_nj1} & \cdots \\ a_{t_n12} & a_{t_n22} & \cdots & a_{t_nj2} & \cdots \\ a_{t_n13} & a_{t_n23} & \cdots & a_{t_nj3} & \cdots \\ a_{t_nj14} & a_{t_n24} & \cdots & a_{t_nj4} & \cdots \\ a_{t_n15} & a_{t_n25} & \cdots & a_{t_nj5} & \cdots \end{pmatrix},$$

$$z_{ji}^n = \left(1/N_{t_n} \right) \bullet \left(p_{t_nj1}{}', p_{t_nj2}{}', p_{t_nj3}{}', p_{t_nj4}{}', p_{t_nj5}{}' \right) \begin{pmatrix} a_{t_nj1} \\ a_{t_nj2} \\ a_{t_nj3} \\ a_{t_nj4} \\ a_{t_nj5} \end{pmatrix}. \tag{7}$$

Among them, the total number of comments is N_{t_n}, the number of medium comments is N_{t_n1}, and the number of bad comments is N_{t_n2}.

It has the following relationship with the number of "bad reviews" $a_{t_k/1}\,(j = 1, 2, 3, \ldots)$, the number of "moderate reviews" $a_{t_{s/2}}$, the number of "general reviews" $a_{t_{t/3}}$, the number of "good reviews" $a_{t_n/4}$, and the number of "very good reviews" a_{t_hj5} obtained by the cluster analysis method

$$\begin{aligned} a_{t_nj3} &= N_{t_n1} + N_{t_n2} - a_{t_n1} - a_{t_n2} \\ a_{t_nj4} &= N_{t_n} - \left(N_{t_n1} + N_{t_n2} \right) - a_{t_nj5}. \end{aligned} \tag{8}$$

2.1.2. The Text Keyword Is a Numerical Value. In the actual situation, the keywords of some factors are specific numerical values. For example, the expression form of the speed of ticket issuance and the speed of ticket issuance in online comments generally emphasize the speed at a specific time, which is based on the expression form of the comment language. Taking this as an example, we might as well set the keyword (that is, the specific value in the comment) of the ticket issuance time and ticket issuance speed in the t_n time period as $c_{t_n}{}'$, and the expected value as $c_{t_n}{}''$. The following relationship exists:

$$\begin{aligned} & c'_{t_nj} \in \left[0, d_{t_hj} \right) \text{and } c_{t_nj}{}' \geq \frac{c''_{t_nj}}{2}, \\ & \frac{2d_{t_5j}}{5} \leq 2c_{t_nj}{}'' \leq d_{t_nj}. \end{aligned} \tag{9}$$

Then, there is

$$z_{ji}^n = 5\left(0.75 - \frac{c_{t_nj}{}' - c_{t_nj}{}''}{d_{t_nj}} \right). \tag{10}$$

2.2. Comprehensive Evaluation of Film Reviews. The paper adopts the fuzzy comprehensive evaluation method. It mainly evaluates similar movies in different theaters through online review data sources. The specific steps are as follows:

(1) Setting the factor set

Through the keyword clustering analysis method (see Section 3), $\alpha = 0.75$ is taken, and the factor set of the content described by the keyword, that is, the content described by the keyword, is $\{s_1, s_2, s_3, s_4\}$. It represents the "service attitude," "ticket issuance speed," "ticket issuance speed," and "film quality" of the film theater, respectively.

(2) Determining the evaluation set

In the nth film evaluation, regarding the factor $s_j\,(j = 1, 2, 3, 4)$, the evaluation set V = {NB, NS, ZO, PS,PB} of the film represents the nth satisfaction evaluation "very dissatisfied," "dissatisfied," "average," "satisfied," and "very satisfied." Its corresponding fuzzy set membership function set $V' = \left\{ \omega_{j1}^n, \omega_{j2}^n, \omega_{j3}^n, \omega_{j4}^n, \omega_{j5}^n \right\}$.

(3) Establishing a weight set

The weight set Q = {0.25, 0.2, 0.15, 0.4} is set.

(4) Obtaining the judgment matrix

Similar films in different theaters are evaluated by analyzing online review data. That is, similar films u_{ni} in each film theater on Ω_n are evaluated. According to the factor $s_j\,(j = 1, 2, 3, 4)$ in the factor set, the online comment keywords are analyzed, and then the membership degree r_{ni}^{jk} of the fuzzy set z_{ni}^j on the evaluation set ω_{jk}^n is calculated by X.

We assume that $\underline{r_{ni}^j} = \left(\underline{r_{ni}^{j1}} \;\; \underline{r_{ni}^{j2}} \;\; \underline{r_{ni}^{j3}} \;\; \underline{r_{ni}^{j4}} \;\; \underline{r_{ni}^{j5}} \right)$, then the judgment matrix R_{ni} can be expressed as

$$R_{ni} = \begin{Bmatrix} r_{ni}^{11} & r_{ni}^{12} & r_{ni}^{13} & r_{ni}^{14} & r_{ni}^{15} \\ r_{ni}^{21} & r_{ni}^{22} & r_{ni}^{23} & r_{ni}^{24} & r_{ni}^{25} \\ r_{ni}^{31} & r_{ni}^{32} & r_{ni}^{33} & r_{ni}^{34} & r_{ni}^{35} \\ r_{ni}^{41} & r_{ni}^{42} & r_{ni}^{43} & r_{ni}^{44} & r_{ni}^{45} \end{Bmatrix}. \tag{11}$$

We assume that $\overline{r_{ni}^j} = \left(\overline{r_{ni}^{ji}} \;\; \overline{r_{ni}^{j2}} \;\; \overline{r_{ni}^{j3}} \;\; \overline{r_{ni}^{j4}} \;\; \overline{r_{ni}^{j5}} \right)$, then the judgment matrix $\overline{R_{ni}}$ can be expressed as

$$\overline{R_{ni}} = \begin{Bmatrix} \overline{r_{ni}^{11}} & \overline{r_{ni}^{12}} & \overline{r_{ni}^{13}} & \overline{r_{ni}^{14}} & \overline{r_{ni}^{15}} \\ \overline{r_{ni}^{21}} & \overline{r_{ni}^{22}} & \overline{r_{ni}^{23}} & \overline{r_{ni}^{24}} & \overline{r_{ni}^{25}} \\ \overline{r_{ni}^{31}} & \overline{r_{ni}^{32}} & \overline{r_{ni}^{33}} & \overline{r_{ni}^{33}} & \overline{r_{ni}^{35}} \\ \overline{r_{ni}^{41}} & \overline{r_{ni}^{42}} & \overline{r_{ni}^{43}} & \overline{r_{ni}^{44}} & \overline{r_{ni}^{45}} \end{Bmatrix}. \tag{12}$$

(5) Comprehensive evaluation.

We assume that $s = s_1 \otimes s_2 \otimes s_3 \otimes s_4$ means that the four factor sets are integrated.

Moreover, the nth comprehensive evaluation conclusions "bad review," "moderate review," "general review," "good review," and "excellent review" are expressed as $C_1^n, C_2^n, C_3^n, C_4^n, C_5^n$, respectively. C_k^n and ω_{jk}^n have the same membership function; here, $k = 1,2,3,4,5$.

We assume $B_{ni} = Q^{\circ} R_{ni} = \left(b_{ni}^1 \ b_{ni}^2 \ b_{ni}^3 \ b_{ni}^4 \ b_{ni}^5 \right)$, that is,

$$\begin{aligned} \underline{B_{ni}} &= Q^{\circ} \underline{R_{ni}} = \left(b_{ni}^1 \ \underline{b_{ni}^2} \ \underline{b_{ni}^3} \ \underline{b_{ni}^4} \ \underline{b_{ni}^5} \right), \\ \overline{B_{ni}} &= Q^{\circ} \overline{R_{ni}} = \left(b_{ni}^1 \ \overline{b_{ni}^2} \ \overline{b_{ni}^3} \ \overline{b_{ni}^4} \ \overline{b_{ni}^5} \right). \end{aligned} \tag{13}$$

Here, $b_{ni}^k = [\underline{b_{ni}^k}, \overline{b_{ni}^k}] \ (k = 1, 2, 3, 4, 5)$ means that u_{ni} comprehensively evaluates the membership degree of C_k^n. Furthermore, there are

$$C_k^n = \sum_{i=1}^{m_n} \frac{1/b_{ni}^k}{u_{ni}} = \sum_{i=1}^{m_n} \frac{1/\left[\underline{b_{ni}^k} \overline{b_{ni}^k} \right]}{u_{ni}}. \tag{14}$$

In this way, the language dynamics trajectory of the comprehensive evaluation value of the similar film theaters is formed as follows:

$$C_k^1, C_k^2, \cdots, C_k^n, \cdots. \tag{15}$$

2.3. The Linguistic Dynamics Trajectory of Film Criticism in the Time-Varying Universe. Through the analysis of online reviews, further improvement measures can be implemented for theaters to achieve closed-loop control of theater film transactions. Cinema improvement measures are to exert influence in a planned and step-by-step manner after analyzing the current situation and problems of the cinema, so as to make the operation change towards the desired goal.

In the paper, IT2FS is used to describe the cinema improvement measures. Due to the different factors that affect the current situation of the theater, the theater film is not the same as the problem now. Different factors have different degrees of influence, and the corresponding improvement measures will also be different. Through the keyword cluster analysis method, $\alpha = 0.75$ is taken. The content of the keyword description, that is, the factor set of the content described by the keyword, is $\{s_1, s_2, s_3, s_4\}$, and it represents the "service attitude," "ticketing speed," "ticketing speed," and "film quality" of the film theater, respectively. Regarding the factor $s_j \ (j = 1, 2, 3, 4)$, the corresponding fuzzy set membership function set of the film evaluation set is $\left\{ \omega_{j1}^n, \omega_{j2}^n, \omega_{j3}^n, \omega_{j4}^n, \omega_{j5}^n \right\} = \{\text{NB}, \text{NS}, \text{ZO}, \text{PS}, \text{PB}\}$. The corresponding improvement measures are $l_j \ (l_j = 4)$ categories, and the membership function set of the strength of each category is $\{v_{1i}^n, v_{2i}^n, v_{3i}^n\} = \{\text{Bad}, \text{Normal}, \text{Good}\}$. In the actual cinema film reviews, the data collection time of online reviews is generally the sales in the last 30 days. In the paper, the length of the data collection period is 30 days. Taking time t as a cycle, the fuzzy rules are established through the experience of experts and the statistical survey of online actual data as follows:

R_1: if $s_1(n)$ is $\omega_{1m_{k_1}}, U_1$ is v_{jl_1}, then $s_1(n+1)$ is $\omega_{1m_{81}}$;
R_2: if $s_2(n)$ is $\omega_{2m_{k_2}}, U_2$ is $v_{jl_{r2}}$, then $S_2(n+1)$ is $\omega_{2m_{g2}}$;
R_3: if $s_3(n)$ is $\omega_{3m_{k3}}, U_2$ is $v_{jl_{r3}}$, then $s_3(n+1)$ is $\omega_{3m_{83}}$;
R_4: if $s_4(n)$ is $\omega_{4m_{k4}}, U_2$ is $v_{jl_{r4}}$, then $s_4(n+1)$ is $\omega_{4m_{g4}}$.

Among them, $m_{k_j}, m_{g_j} \in [1, 5], l_{r_j} \in [1, 3]$, and they are all positive integers. R_j is a fuzzy rule defined on Ω_j, including 15 fuzzy subrules, which are defined as follows:

$R_j(1,1)$: if $s_J(n)$ is NB, U_j is bad, then $s_j(n+1)$ is NB;
$R_j(1,2)$: if $s_j(n)$ is NB, U_j is normal, then $s_j(n+1)$ is NS;
$R_j(1,3)$: if $s_j(n)$ is NB, U_j is good, then $s_j(n+1)$ is ZO;
$R_j(2,1)$: if $s_J(n)$ is NS, U_j is bad, then $s_j(n+1)$ is NS;
$R_j(2,2)$: if $s_j(n)$ is NS, U_j is normal, then $s_j(n+1)$ is ZO;
$R_j(2,3)$: if $s_J(n)$ is *NS*, U_j is good, then $s_j(n+1)$ is PS;
$R_j(3,1)$: if $s_j(n)$ is ZO, U_j is bad, then $s_j(n+1)$ is ZO;
$R_j(3,2)$: if $s_j(n)$ is ZO, U_j is normal, then $s_j(n+1)$ is ZO;
$R_j(3,3)$: if $s_J(n)$ is ZO, U_j is good, then $s_j(n+1)$ is PS;
$R_j(4,1)$: if $s_J(n)$ is PS, U_j is bad, then $s_j(n+1)$ is PS;
$R_j(4,2)$: if $s_j(n)$ is PS, U_j is normal, then $s_j(n+1)$ is PS;
$R_j(4,3)$: if $s_J(n)$ is PS, U_j is good, then $s_j(n+1)$ is PB;
$R_j(5,1)$: if $s_j(n)$ is PB, U_j is bad, then $s_j(n+1)$ is PB;
$R_j(5,2)$: if $s_J(n)$ is PB, U_j is normal, then $s_j(n+1)$ is PB;
$R_j(5,3)$: if $s_J(n)$ is PB, U_j is good, then $s_j(n+1)$ is PB.

The initial word $\omega_j(0)$ is calculated through comprehensive evaluation, and the activated fuzzy rule R_j is determined. The improvement goal ω_j^* is given, then there are

$$\lambda^* \left(m_{k_j}, 0 \right) = \omega_j^* \wedge \omega_{jm_{gj}}. \tag{16}$$

Furthermore, there are

$$\lambda^* \left(m_{k_j}, 0 \right) = \max \left\{ \lambda^*(1,0), \lambda^*(2,0) \ldots, \lambda^* \left(m_{k_j}, 0 \right) \right\}. \tag{17}$$

According to the above formula, it can be calculated that v_{jl_j} is the most suitable improvement measure.

In order to facilitate the calculation, the paper takes improvement measures from four aspects: service, ticket issuance, film source, and work efficiency. We assume that $\Omega_n = \{u_{n1}, u_{n2}, \ldots, u_{nm}\}$ is a discrete time-varying universe, where $u_{ni} \ (\text{i} = 1, 2, \ldots)$. At the same time, we assume that $U = \{U_1, U_2, U_3, U_4\}$ is a discrete universe of discourse and use three fuzzy sets $v_{1i}^n, v_{2i}^n, v_{3i}^n$ to cover the universe of universe U, which, respectively, represent the effect of the nth improvement measures "poor," "average," and "good." $v_l^n(u_{ni}) \ (l = 1, 2, 3)$ represents the degree to which the evaluation scores of the corresponding improvement measures taken for the similar films in the ith theater belong to the fuzzy set X_l^n after the nth evaluation. Corresponding to

"poor," "general," and "good" specific measures, the FOU of its membership function is shown in Figure 2.

After applying the division, its membership function can be expressed as

$$v_l^n = \int_{u\in\Omega}\int_{x\in L_{ui}^{'n}} \frac{\mu_{X_i^n}(u,x')}{(u,x')}, L_{ul}^{'n} \subseteq [0,1] \text{and} \mu_{v_l^n}(u,x') = 1. \tag{18}$$

Among them, there are

$$L_{ul}^{'n} = \left[\underline{g_l^{'n}}, \overline{g_l^{'n}}\right],$$
$$FOU\left(v_l^n\right) = \underset{u\in\Omega}{\cup} u \times L_{ul}^{'n}. \tag{19}$$

Among them, $L_{ul}^{'n} = [\underline{g_l^{'n}}, \overline{g_l^{'n}}]$, the lower boundaries $\underline{g_l^{'n}}, \overline{g_l^{'n}}$ can be expressed as

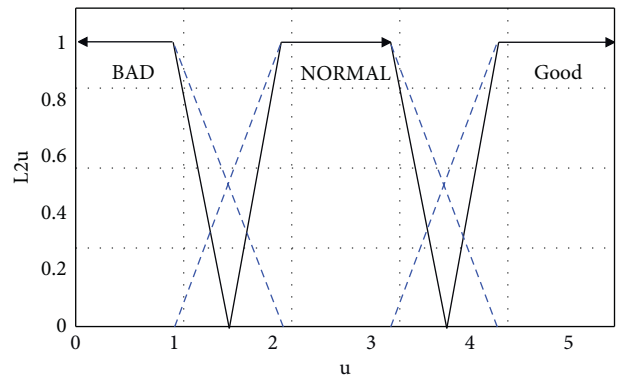

Figure 2: FOU of the membership function of the improvement measures.

$$\underline{g_1^{'n}} = \begin{cases} 1 & 0 \le z_{li}^{'n} < 1 \\ \dfrac{1.5 - z_{li}^{'n}}{0.5} & 1 \le z_{li}^{'n} < 1.5 \\ 0 & 1.5 \le z_{li}^{'n} \le 5 \end{cases}, \underline{g_3^{'n}} = \begin{cases} 0 & 0 \le z_{li}^{'n} < 3.5 \\ \dfrac{z_{li}^{'n} - 3.5}{0.5} & 3.5 \le z_{li}^{'n} < 4 \\ 1 & 4 \le z_{li}^{'n} \le 5 \end{cases},$$

$$\underline{g_2^{'n}} = \begin{cases} 0, & 0 \le z_{li}^{'n} \le 1.5 \\ \dfrac{z_{li}^{'n} - 1.5}{0.5}, & 1.5 < z_{li}^{'n} \le 2 \\ 1, & 2 < z_{li}^{'n} \le 3, \end{cases} \overline{g_2^{'n}} = \begin{cases} 0, & 0 \le z_{li}^{'n} \le 1 \\ z_{li}^{'n} - 1, & 1 < z_{li}^{'n} \le 2 \\ 1, & 2 < z_{li}^{'n} \le 3 \\ \dfrac{3.5 - z_{ji}^n}{0.5} & 3 < z_{li}^{'n} \le 3.5 \\ 0, & 3.5 < z_{li}^{'n} \le 5 \end{cases} \begin{cases} 4 - z_{ji}^n, & 3 < z_{li}^{'n} \le 4 \\ 0, & 3.5 < z_{li}^{'n} \le 5 \end{cases}, \tag{20}$$

$$\overline{g_1^{'n}} = \begin{cases} 1 & 0 \le z_{li}^{'n} < 1 \\ 2 - z_{li}^{'n} & 1 \le z_{li}^{'n} < 2 \\ 0 & 2 \le z_{li}^{'n} \le 5 \end{cases}, \overline{g_3^{'n}} = \begin{cases} 0 & 0 \le z_{li}^{'n} < 3 \\ z_{li}^{'n} - 3 & 3 \le z_{li}^{'n} < 4 \\ 1 & 4 \le z_{li}^{'n} \le 5 \end{cases}$$

Among them, $z_{li}^{'n}$ is the improvement intensity value indicating the corresponding improvement measures taken for the similar films in the ith theater after the nth evaluation.

2.4. The Language Dynamics System of IT2FS under the Time-Varying Universe. It is difficult for conventional mathematical models to study the complex system of human direct participation in film analysis. People can establish fuzzy rules based on actual experience and relevant knowledge, and then use fuzzy reasoning to analyze the system dynamically. Therefore, a language dynamics system of interval type-two fuzzy sets (IT2FSs) in the time-varying universe is proposed. We assume that $R_1, R_2, \cdots, R_n, \cdots$ are dynamic fuzzy rules defined on the universes $\Omega_1, \Omega_2, \cdots, \Omega_n, \cdots$ and n is a natural number, then there are

$$R = \bigcup_{n=1}^{\infty} R_n. \tag{21}$$

For the initial state word $\omega_1 \in \mathfrak{R}(\Omega_1)$ and the initial input word $v_1 \in \mathfrak{R}(U_1)$, under the action of fuzzy rules, the language dynamics trajectory is

$$\omega_1, \omega_2, \cdots, \omega_n, \cdots. \tag{22}$$

Among them, there are

$$\begin{aligned} \omega_{n+1} &= R(\omega_n, v_n) = R^n(\omega_1, v_1), \\ &\omega_n \in \mathfrak{R}(\Omega_n), v_n \in \mathfrak{R}(U_n), \\ R^n &= R_n \circ R_{n-1} \circ \cdots \circ R_1. \end{aligned} \tag{23}$$

The corresponding language dynamics trajectory can also be in the following form:

$$\omega_1, R^1(\omega_1, v_1), \cdots, R^n(\omega_1, v_1), \cdots. \tag{24}$$

For complex systems such as film analysis, the object is generally a single individual. However, individual characteristics factor papers are available in a continuous IT2FS study. Based on the continuous IT2FS, the calculation steps of the language dynamics trajectory of IT2FS in the time-varying universe are as follows:

(1) The domain of discourse of the target object is determined. According to the specific research object, the universe of discourse $\Omega_t, t = 1, 2, 3, \cdots$ at time t is determined.

We assume $\Omega_t = \{u_1, u_2, \cdots, u_n\}$. If Ω_t is an incremental discrete time-varying universe, then $\Omega_{t+1} = \{u_1, u_2, \cdots, u_n, u_{n+1} \cdots u_{n+i}\}$. If Ω_t is a decreasing discrete time-varying universe, then $\Omega_{t+1} = \{u_1, u_2, \cdots, u_{n-i-1}, u_{n-i}\}$. If Ω_t is a wave-type discrete time-varying universe, then $\Omega_{t+1} = \{a_1, a_2, \cdots, a_{n-i-1}, a_{n-i}, a_{n+1} \cdots a_{n+i}\}, i \geq 1$.

(2) The feature set of each element in the universe of discourse is determined. According to the characteristics of the domain element, the possible characteristics of the element are determined and covered with the set base words.

(3) The algorithm evaluates the object, and the paper adopts the fuzzy comprehensive evaluation. Elements in different domains of discourse at different times have different feature sets, and the same element has different membership degrees for different features at different times. Therefore, with the change of time t, the language dynamics trajectory of the system can be obtained.

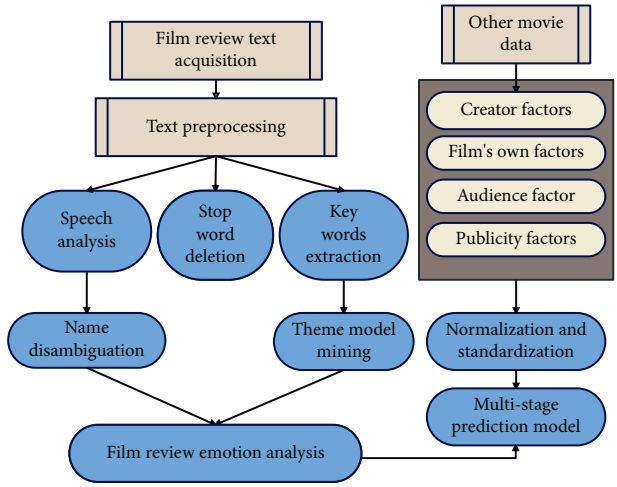

Figure 3: Overall framework of multistage box office forecasting.

However, in intelligent control, the author needs to set the expected target according to the control object, and then formulate interference or improvement measures. A closed-loop control is formed to obtain the expected language dynamics trajectory and complete the control task.

3. Research on Visualization of Film Review Data Based on Cluster Analysis Technology

The contribution value of an actor can be evaluated by using Weibo data, star fan data, and published content related to the film release cycle. Sentiment features are given by the analysis of film reviews. Usually, it is necessary to identify the film review for a certain main creator from a large number of comments, then perform statistical analysis on the positive and negative emotions of the film review, and combine the syntactic analysis and fan data to give the influence of the main creator of the change. Figure 3 shows the overall framework of multistage box office prediction.

The main modules of the web comment crawler are web page information collection module, web page storage module, and comment storage module, as shown in Figure 4(a). The crawler crawls the web page through the seed URL, extracts the qualified URLs in the web page, and

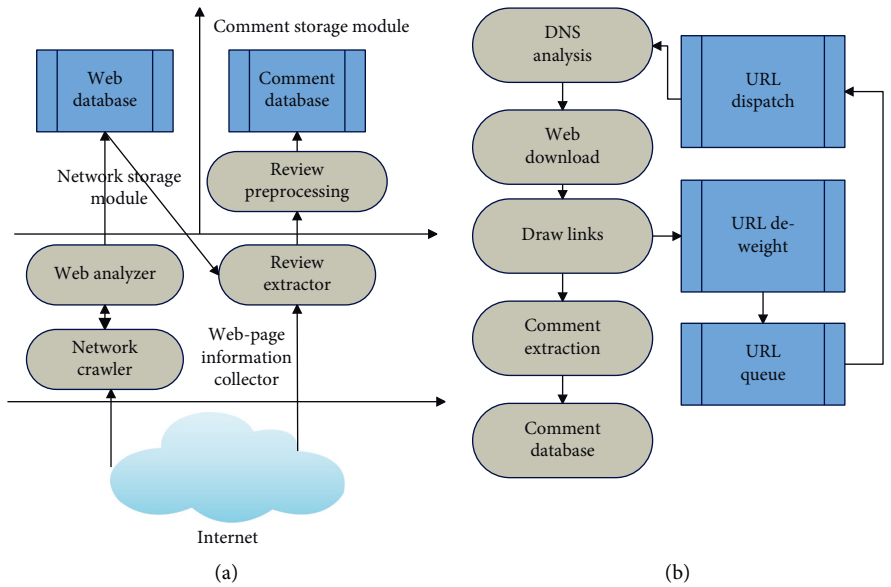

Figure 4: Crawler framework of film review data: (a) architecture of the review crawler and (b) flowchart of the review crawler.

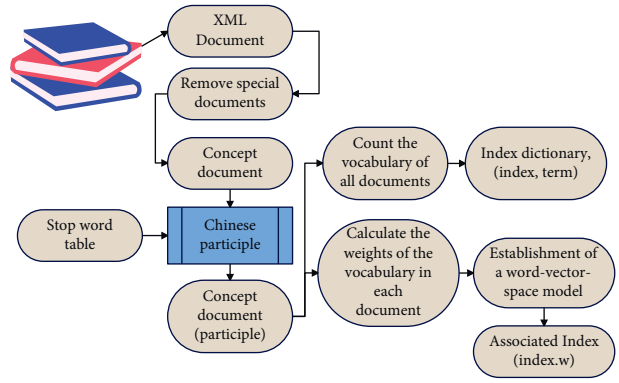

Figure 5: The establishment of the vector space model of words.

adds them to the URL queue to prepare for crawling in the next cycle. Then, it extracts the network comments in the web page through the web page information collection module. This paper uses multithreading to implement web crawler, realizes URL deduplication through bloom filter, and crawls needed comment information. The specific execution steps are as follows, and the flowchart is shown in Figure 4(b).

This paper represents a vocabulary with a set of network natural concepts, encoding reviews using a large amount of highly advanced human knowledge, which is defined by people themselves and can be easily interpreted. Through continuous development, film reviews have grown in breadth and depth over time. The establishment of the vector space model of words is shown in Figure 5. This method makes up for the problem of low accuracy in calculating similarity caused by short vocabulary and high dimensionality and few values.

Emotional dictionaries are generally not perfect. Moreover, when dealing with different text sentiment analysis tasks, the coverage of sentiment words is obviously not enough. Analyzing text information in the film industry requires building a sentiment dictionary that fits the product, and analyzing the film industry also requires building a sentiment dictionary suitable for film review analysis. Therefore, according to the actual research situation, it is necessary to mine new emotional words and add them to the emotional dictionary used for research. The process framework of constructing emotion dictionary in the film field is shown in Figure 6.

After data preprocessing, the LDA topic model is used for topic modeling. In order to analyze the sentiment tendency from a topic that is more in line with the market perspective, through the custom topic and the initialized attribute word, the distance between the Word2Vec word vectors is used, and the attributes with a correlation greater than 0.8 are selected for topic classification. These two subject-attribute word sets are fused to construct a subject-attribute word set, as shown in Figure 7.

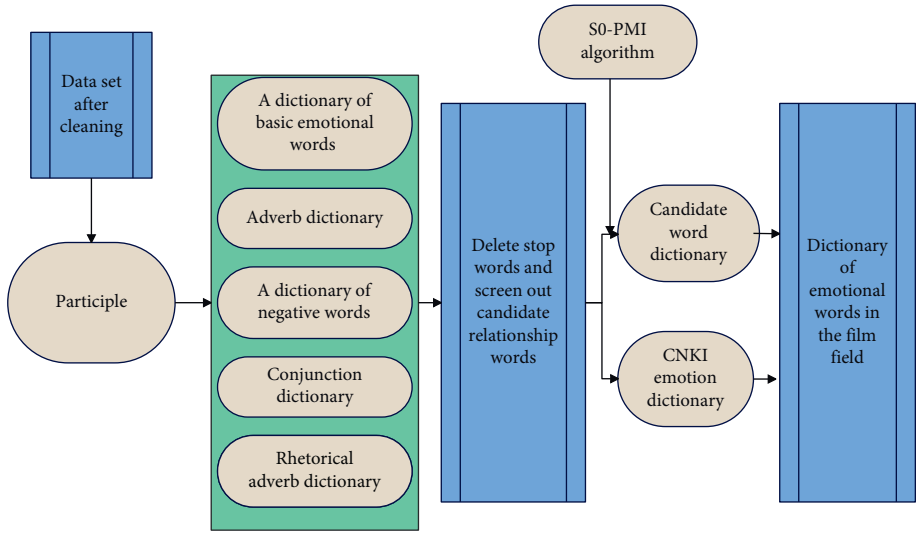

Figure 6: Construction framework of emotional dictionary in the film field.

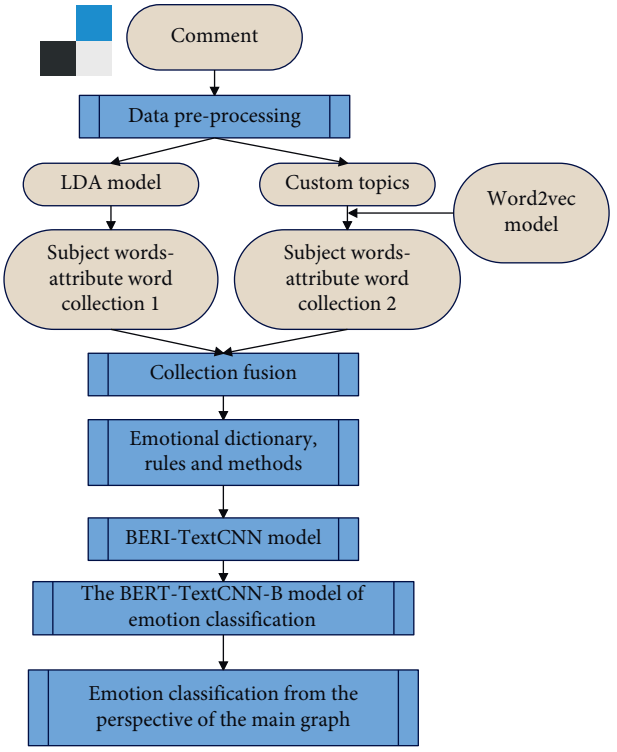

Figure 7: A fine-grained sentiment analysis framework for film reviews.

When reducing the dimension of film reviews, the vector dimension obtained after training for each film is 400 dimensions, which is too high. Therefore, when the traditional t-SNE algorithm is used for dimensionality reduction, the result shown in Figure 8(a) is obtained. There are problems that the clustering results are not clear, they do not meet the clustering principle of "high cohesion and low coupling," and the significant effect between the clustering results is not ideal.

The PCA predimension reduction process is introduced, and the 400-dimensional high-dimensional vector is first reduced to 50 dimensions, and then the t-SNE dimensionality reduction algorithm is used to solve the problem of poor cluster visualization. The clustering results after introducing PCA predimension reduction are shown in Figure 8(b).

The t-SNE method with the introduction of PCA predimensionality reduction is used to visualize the results of

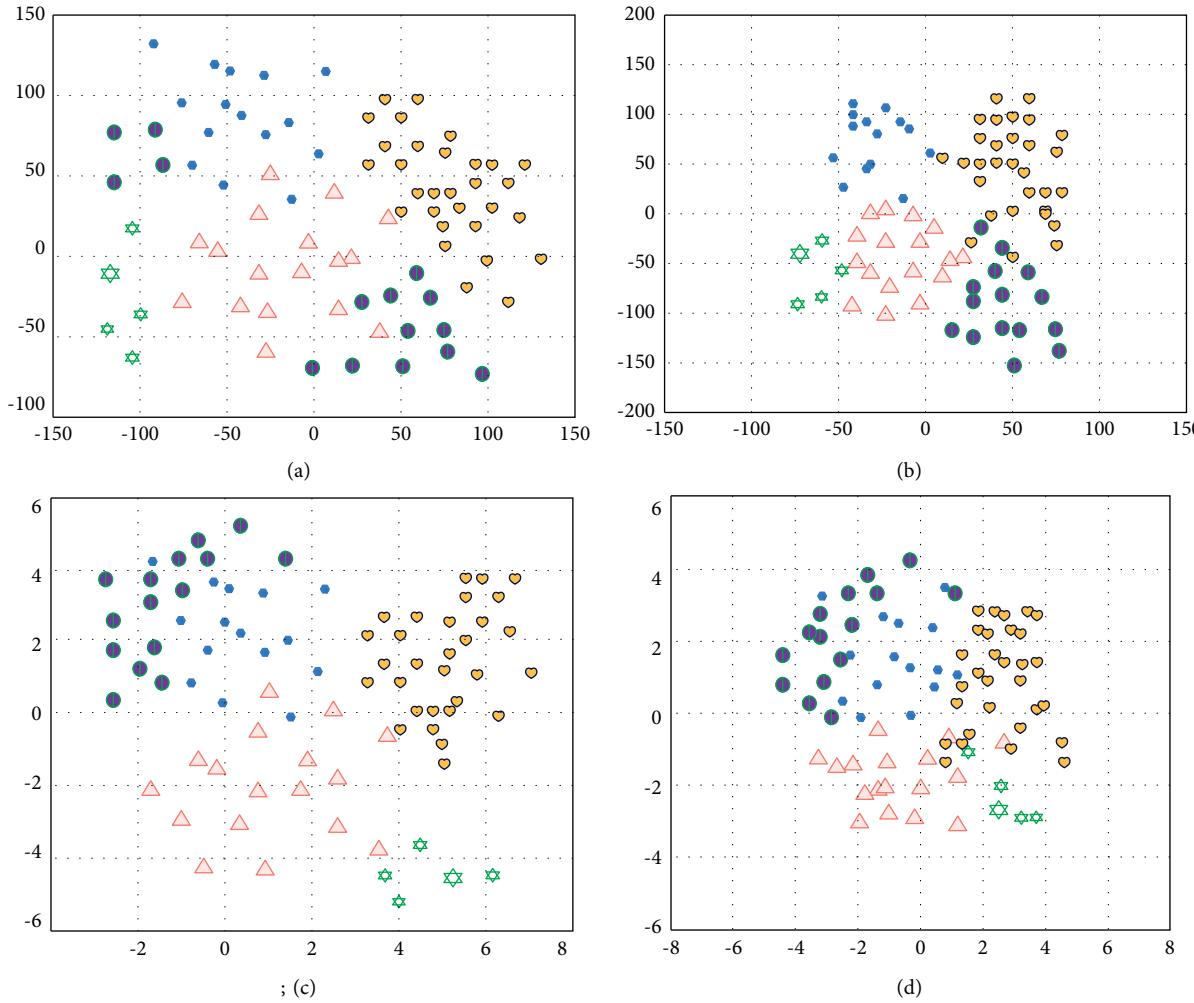

Figure 8: Film review clustering effect: (a) *k*-means clustering results; (b) *k*-means clustering results after introducing PCA predimension reduction; (c)clustering results of the chasing clustering algorithm without PCA prereduction; (d)clustering results of the chasing clustering algorithm after introducing PCA predimension reduction.

the chasing clustering algorithm for dimensionality reduction. Comparing Figures 8(c) and 8(d), we can see that the clustering results in Figure 8(d) have better significant effect and meet the clustering requirements of "high cohesion and low coupling."

It can be seen from the above research that cluster analysis technology can play a certain role in the visualization analysis of film review data.

4. Conclusion

Limited by the bottleneck of natural language understanding technology, the industry's use of film review data is still very limited. At present, no film platform or research paper can effectively mine the hidden value in film review data. The film review is the most direct evaluation of the film by the audience, and it can reflect the audience's viewing experience in the most real and detailed way. Moreover, mining the value of film review data is of great significance for us to fully understand the audience experience and scientifically guide the development and progress of the film industry. This paper combines the clustering analysis technology to construct a research system of film review cluster visualization, which can improve the visualization effect of film reviews and promote the film to be better understood by people. The experimental research shows that the cluster analysis technology can play a certain role in the visual analysis of film review data.

Data Availability

The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the Department of Digital Image in Sangmyung University.

References

[1] E. Castano, "Art films foster theory of mind," *Humanities and Social Sciences Communications*, vol. 8, no. 1, pp. 1–10, 2021.

[2] L. Jayyusi, "Hollywood's transnational imaginaries: colonial agency and vision from Indiana Jones to World War Z," *Continuum*, vol. 32, no. 3, pp. 355–369, 2018.

[3] S. Kumar, K. De, and P. P. Roy, "Movie recommendation system using sentiment analysis from microblogging data," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 4, pp. 915–923, 2020.

[4] E. Taiebi Javid, M. Nazari, and M. R. Ghaeli, "Social media and e-commerce: a scientometrics analysis," *International Journal of Data and Network Science*, vol. 3, no. 3, pp. 269–290, 2019.

[5] B. Zou, M. Nurudeen, C. Zhu, Z. Zhang, R. Zhao, and L. Wang, "A neuro-fuzzy crime prediction model based on video analysis," *Chinese Journal of Electronics*, vol. 27, no. 5, pp. 968–975, 2018.

[6] S. S. Sundar, "Rise of machine agency: a framework for studying the psychology of human–AI interaction (HAII)," *Journal of Computer-Mediated Communication*, vol. 25, no. 1, pp. 74–88, 2020.

[7] S. C. Chang, "Market size matters? An approach to illustrate the market preference of Hong Kong-mainland China co-production cinema," *Journal of International Communication*, vol. 26, no. 1, pp. 125–149, 2020.

[8] A. Kaplan and M. Haenlein, "Rulers of the world, unite! the challenges and opportunities of artificial intelligence," *Business Horizons*, vol. 63, no. 1, pp. 37–50, 2020.

[9] B. Kuklick, "Fascism comes to America," *International Journal for History, Culture and Modernity*, vol. 6, no. 1, pp. 1–18, 2018.

[10] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: feature selection for opinion classification in web forums," *ACM Transactions on Information Systems*, vol. 26, no. 3, pp. 1–34, 2008.

[11] R. Piryani, V. Gupta, and V. K. Singh, "Movie Prism: a novel system for aspect level sentiment profiling of movies," *Journal of Intelligent and Fuzzy Systems*, vol. 32, no. 5, pp. 3297–3311, 2017.

[12] L. Pang, "Mediating the ethics of technology: hollywood and movie piracy," *Culture, Theory and Critique*, vol. 45, no. 1, pp. 19–32, 2004.

[13] J. H. Shon, Y. G. Kim, and S. J. Yim, "Classifying movies based on audience perceptions: MTI framework and box office performance," *The Journal of Media Economics*, vol. 27, no. 2, pp. 79–106, 2014.

[14] P. Bosc, D. Dubois, and H. Prade, "Fuzzy functional dependencies and redundancy elimination," *Journal of the American Society for Information Science*, vol. 49, no. 3, pp. 217–235, 1998.

[15] S. Agrawal, R. K. Singh, and Q. Murtaza, "Prioritizing critical success factors for reverse logistics implementation using fuzzy-TOPSIS methodology," *Journal of Industrial Engineering International*, vol. 12, no. 1, pp. 15–27, 2016.

[16] C. Porcel, A. Tejeda-Lorente, M. A. Martínez, and E. Herrera-Viedma, "A hybrid recommender system for the selective dissemination of research resources in a technology transfer office," *Information Sciences*, vol. 184, no. 1, pp. 1–19, 2012.