A Realization Framework of International Image Prediction System for Film and TV Works

Shan Liu, Xiaoqing Wu, Mingyue Zheng, Xinqiao Cheng, Ruijing Fang, Pingru Long
Faculty of Science and Engineering,
Communication University of China
Beijing 100024, China

Abstract—**International image is the window of a country to the world. This paper introduces how we use technologies such as reptile technology, natural language processing, clustering, and data visualization to collect relevant information about film and television works and analyze the internationally spread image of China's film and television works. In addition, an emotion dictionary is established, the emotion index is calculated, and the positiveness of the image is evaluated. Finally, a prediction system was established, and the two functions of the prediction of communication image and recommendation of communication strategy are realized through preset algorithms. The system can be applied to the work of cultural transmission to help build a national image consistent with China's long-standing culture and prosperity.**

Keywords- system development; natural language processing; sentiment analysis; image analysis

I. INTRODUCTION

The concept of "national image" was mostly recognized by Wenhu Guan, that is, the national image is a complex, which is the general evaluation and recognition of the external and internal public, behavior, activities, and results of the country. A national image has great influence and cohesion, which is the embodiment of the overall strength of a country [1]. The national image plays an important role in reflecting national identity, arousing cohesion, exporting values, and increasing influence. In the context of globalization, countries gradually focus on the construction of soft power, hoping to enhance their voice and influence in the world. As a display of a country's value appeal and cultural output, national image is one of the most direct manifestations of national soft power.

People's understanding of a country is often not directly obtained from first-hand information, and mainly through TV, radio, newspapers, magazines, movies, and even networks to obtain secondary data, they learn from these data, with their own cultural and ideological processing digestion, form their own understanding, and complete the country's national image building [2]. The most effective way to build a national image is through the media, which brings convenience to communication and will not be restricted by national regions. On the one hand, the film and television works carried by the media are easy to accept by people through their audio-visual combination, and on the other hand, the image spread is more three-dimensional. Building national images through film and television is an effective way to export culture. However, how to meet the demand for "telling Chinese stories well" proposed by general secretary Xi Jinping, and how to increase the spread of film and television works to give a positive "voice of China" is a major problem now.

At present, the Chinese film and television works disseminated overseas are mostly untrue and not positive in shaping the country's image. American political theorist Lippmann has proposed the theory of "mimicry environment" [3], that is, the environment that the media provided to people after selecting, processing, and restructuring events or information. Because this processing, selection, and structuring took place out of sight (inside the media), people were often unaware of this and tended to view the mimicry environment as the objective environment itself." For overseas groups, what they get was not the real reflection of the objective environment, but the mimicry environment selected by the media, which was the "Chinese image" of media organizations with a certain subjective consciousness.

Now many film and television producers are blindly seeking foreign markets, ignoring the deviations caused by cultural differences at home and abroad on the values conveyed by film and television works, resulting in the spread of inappropriate national images. Most domestic scholars' conclusions still remained subjective in analyzing and judging the communication images and only aimed at a specific work or type; while this paper proposed a scientific data method to analyze and predict the communication image of film and television works, innovatively starting from the relationship between the characteristics of the work and the image, established a model and built a system to solve the problem that their conclusions had too many subjective factors which may lead to deviations and was unfit with other film and television works.

II. THEORTICAL ALGORITHM

A. Comment Clustering Analysis

1) cosine distance calculation of similarity

For high-dimensional numerical data, literature [4] had demonstrated that Euclidean distance was better adapted as a similarity measure of k-means, so the k-means algorithm generally adopted Euclidean distance. For text vectors, literature [5] proposed that cosine similarity was better than Euclidean distance. Literature [6] had pointed out that it paid attention to the absolute value of the numerical difference in dimensions, as Euclidean distance measured distance by the size of the interval length between vectors; while cosine similarity used the cosine

978-0-7381-0945-1/20/$31.00 ©2020 IEEE

value of the included angle to describe the similarity, so it paid more attention to the difference in relative levels between dimensions. In the judgment of text similarity, whether the words appeared simultaneously or not in the same dimension was an important index to judge its similarity, while the number of words appearing or the difference of values in the same dimension was relatively less important. Therefore, the cosine of the angle between vectors could better describe the similarity between text vectors, and the cosine similarity was usually used as the similarity measure for text clustering.

Since the data objects in this project were mainly text information, the paper used cosine distance to calculate similarity.

2) K-center clustering algorithm

The K-center point algorithm is a typical partition-based clustering algorithm, which uses the center point as the clustering center of each cluster. Compared with the k-means algorithm, this method can effectively eliminate the influence of outliers on the clustering effect [7], and improve the clustering accuracy [8].

- The number of clustering K was determined;
- The K center points from all data sets were selected randomly, and each center point corresponded to a cluster;
- The distance from all points to the K center points was calculated, and the points with the shortest distance to the center point were taken as a cluster;
- The point with the smallest absolute error from each point within the cluster was calculated as the new center point;
- Steps 3 and 4 were repeated until the center point of each cluster was no longer changed.

After clustering, the top 5 words were extracted with the highest occurrence frequency for each cluster, and then the image represented by each cluster was summarized. the paper counted the number of data in each cluster and calculated its image proportion. The formula is as follows:

$$\theta(a)=\frac{\varphi(a)}{K} \quad (1)$$

where "a" is the image represented by this cluster, and φ(a) is the number of data in "a" cluster; K is the total number of data; and θ(a) is the proportion of attribute "a".

B. Emotional Analysis System

The paper found emotional words in the text information, scored each emotional word, and finally added up the scores in the sentence.

The scoring rules are as follows:

1) negative rule:

If there are negative words in front of the emotional word, multiply the emotional word score by -1;

2) adverb rule:

According to the degree of adverb expression, multiply by the corresponding degree coefficient;

3) exclamation rule:

Use symbols to emphasize emotions, such as!!! Add 1 to the score;

4) emotion rule:

If the emoji is recognized, add appropriate marks according to the meaning expressed by the emoji.

C. Image Analysis

1) Factors of affecting image

a) Types

The image created by film or television work is often closely related to its type. For example, in the literature [9], in the family ethics drama, the main characters in the drama always had to experience the various miseries and tests in life. Finally, they gained the unanimous approval and respect of society, and realized their eternal value as good people. The same type of film and television works was likely to bring the same impression to the audience because of the similarity of the plots, themes and even the clothing.

b) Area of spreading

Different countries and regions have different cultures. Even if the same film and television works are transmitted in different regions, the images transmitted may be different. People from different countries and regions have different concerns and preferences regarding the film and television works.

c) Year of releasing

The spread of TV dramas is also influenced by the time of broadcast. If there are some related hot events during the release, it will have a great impact on the spread of TV dramas.

d) Number of episodes and Time

The number of episodes and time will affect the viewer's perception of the film and television work.

There are many other factors can influence the images, and all of them, including the above, are features of films and TV Works, expressed in uppercase letters (A, B, C……, M).

2) Calculating the intimacy index – α

The image ratio of an attribute of a work containing a certain feature was averaged to obtain an intimacy index of the attribute for the feature. The intimacy index was calculated as follows:

$$\alpha(A_i, a) = \frac{\sum \theta(A_i, n, a)}{N(A_i \in T)} \quad (2)$$

where T= {{A1, A2, A3...An}, {B1, B2, B3...Bn}, {C1, C2, C3...Cn}, ...}

$\theta(A_i, n, a)$ is the θ value of the attribute "a" in the nth part of the work containing the feature A_i;

$N(A_i \in T)$ is the number of works containing the feature A_i;

$\alpha(A_i, a)$ is the intimacy index of the attribute "a" for the feature A_i.

3) Calculating the comprehensive intimacy index – ψ

a) Inter-feature impact coefficient – k

The formula for calculating the influence coefficient between feature A_i and feature B_j is:

$$k(A_i \cap B_j, a) = 2 * \frac{\alpha(A_i \cap B_j, a)}{\alpha(A_i, a) + \alpha(B_j, a)} \quad (3)$$

And

$$\alpha(A_i \cap B_j, a) = \frac{\sum \theta(A_i \cap B_j, n, a)}{N[(A_i \in T) \cap (B_j \in T)]} \quad (4)$$

If k is approximately equal to 1, it is proved that these two features are almost independent; if k is greater than 1, the two features are mutually promoted; if k is less than 1, the two features are mutually suppressed.

Expanding to M features, the formula for calculating the influence coefficient between M features is:

$$k((A_i \cap B_j \cap \ldots \ldots \cap M_n, a) = M * \alpha(A_i \cap B_j \cap \ldots \ldots \cap M_n, a)/(\alpha(A_i, a) + \alpha(B_j, a) + \cdots + \alpha(M_n, a)) \quad (5)$$

b) Considering the mutual influence coefficient between each factor (consider the case where k is too more or less than 1)

The comprehensive intimacy index formula that considers the effects between features is:

$$\psi(a) = k(A_i \cap B_j \cap C_k \ldots \ldots \cap M_n) * [\alpha(A_i, a) + \alpha(B_j, a) + \alpha(C_k, a) + \cdots + \alpha(M_n, a)] \quad (6)$$

c) Ignoring the influence coefficient (k is nearly equal to 1)

If the idea of controlling variables were adopted, that is, the number of each feature works should be balanced and maximized when selecting works. When calculating the intimacy of a certain feature and an image, the number of works containing the remaining features is as close as possible to eliminate other the effect of features on the results, thereby controlling a single variable. At this point, the influence coefficient between the features could be ignored. At this point, the calculation formula is improved: $\psi(a) = \alpha(A_i, a) + \alpha(B_j, a) + \alpha(C_k, a) + \cdots + \alpha(M_n, a)$ (7)

III. Introduction for System Function

The main content of the paper is to design a visual image prediction system for film and television works. Natural language processing methods were used to extract and comprehensively analyze the images of typical film and television works. Based on it predicts the image that may be conveyed by the film and television works that have not yet been released, and provides suggestions.

The basic design process can be shown as below:

- Data collection and preliminary processing;
- Cluster and Analyze Textual information;
- Analyze the emotional tendency of the text;
- Establish a predictive model;
- Develop an APP;

A. Data collection and preliminary processing

Using reptile technology to collect the dissemination and user comments of film and television works in different countries and regions from foreign video websites, social platforms, film review websites and other platforms. They were translated into one language.

When preprocessing data such as user comments, Python's Chinese participle component, the jieba participle tool, can be used. Here are the steps to deal with user comments:

- A custom dictionary of common nouns (such as place and name) were written in works and reviews.
- The compiled custom dictionary in the jieba thesaurus was loaded. Use the jieba participle tool to extract keywords from sentences based on the TF-IDF value. Or mark the part of speech first, and then extract the keywords with higher TF-IDF value according to the part of speech.
- A dictionary of stop words was written according to Chinese grammar and all stop words from the extracted keywords were deleted.
- By making all the words into a set, each comment becomes a vector array, and all the comments together become a matrix. The column of the matrix is a collection of keywords extracted from all comments. Each row represents a comment. Each row is a vector which comes from a comment and each value of the vector is the TF-IDF weight of the word commented on in the article.

Take Jackie Chan's "Police Story" as an example:

TABLE I. COMMENT LIST

Comment
Jackie Chan 's representative works. In order to create real effects, Jackie Chan is also desperately performing. He suffered a lot of injuries and suffered a lot. Fortunately, the final effect is very good. A high-quality Hong Kong movie is worthwhile. Watch.
Just love to see Maggie Cheung being dragged from the little sheep! ~
Those fight performances!
The most classic movie of Jackie Chan early.
Jackie Chan's early classics, the score is a bit low.
I like the funny Jackie Chan.
Zhang Manyu was really young at that time!
Jackie Chan action classic series.
very classic.
……

The custom dictionaries written included "Jackie Chan", "Zhang Manyu", "Hong Kong" and other words. After deleting the deactivated words, the keywords extracted according to the TF-IDF weight were "Jackie Chan", "Zhang Manyu", "classic", "Early" and so on. Assume that the TF-IDF values of the first keyword "Jackie Chan" in each comment were tf11, tf21, tf31,

tf41, tf51, ..., tf81, and the TF-IDF values of the second keyword "Maggie Cheung" in each comment were tf12, tf22, tf32, tf42, tf52 ... then these comments could form a matrix as follows:

TABLE II. TF-IDF VALUE

Jackie Chan	Zhang Manyu	classic	Early
tf11	tf12	tf13	tf14
tf21	tf22	tf23	tf24
……	……	……	……
tf81	tf82	tf83	tf84
……	……	……	……

B. Text cluster analysis

The cluster analysis of the processed comments could be divided into appropriate clusters, as shown in the figure below:

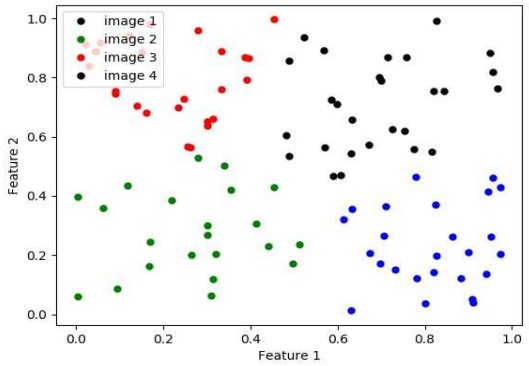

Figure 1. Clustering diagram.

The top five words with the highest frequency in each cluster were extracted and they were summarized as one image represented by the cluster.

Calculate the ratio of the image in this work and store it in the database.

C. Textual Emotional Analysis

The emotional analysis tool is used to analyze the comments under each attribute and obtain an emotional index.

The specific steps are:

- Create four dictionaries representing positive, negative, adverb of degree and negation;
- Find all emotional words in the comments after the participle, and the value of a positive emotional word is 1;
- Estimate whether there is an adverb of degree in front, multiply the degree coefficient according to the degree;
- Estimate whether there is a negative word in front, and if there is a negative word, multiply by -1;
- Add all the values together to obtain a positive emotion index;
- Negative emotion words are the same;
- Emotion index = positive emotion index-negative emotion index.

D. Establish a predictive model

1) Establish an information matrix

a) Image keyword matrix

The clustered words were combined into an image array, and the corresponding value {a b c d ... n} was the proportion of the image in the work. The words obtained by clustering were the Chinese images extracted from the comments. To convert these images into data that can be separately calculated, the proportion of each cluster is calculated and assigned to the array {a b c d ...n}. Eventually, the image array of all works formed an image keyword matrix.

b) Work label matrix

According to the feature information of the work, set each feature category as A, B, C... M. The specific characteristic words in each category were the optional labels of works, which have been set to {A1, A2, A3...An}, {B1, B2, B3...Bn}, {C1, C2, C3, ..., Cn}, ..., {M1, M2, M3, ... Mn}. Specific labels like A1, A2..., An were mutually exclusive. If the work has a label, the corresponding value would be set to 1., and others would be set to 0. Set the work address value to {{A1, A2, A3...An}, {B1, B2, B3...Bn}, {C1, C2, C3, ..., Cn}, ..., {M1, M2, M3... Mn}}. The work label matrix consists of the address values of all works.

e.g. The characteristics of Chinese exporting film and television works are classified as the country, subject matter, release time, and the duration of film and television dramas. Set them to A B C D. The specific labels in the dissemination countries are Japan, South Korea, the United States, and the United Kingdom. These countries are set to A1, A2, A3, A4, the other categories are same to category A. If the work 1 was spread in Japan, we would set A1=1, A2=A3=A4=0 forming an array {1 0 0 0}, and B, C, and D would be the same. The [A B C D] array was formed into the address value of the work 1.

Take " Wolf Warriors Ⅱ" as an example, A1 is suspense, A2 is action, A3 is love, A4 is the plot, B1 is South Korea, B2 is Europe and America, B3 is Japan, B4 is Vietnam, C1 is before 2000, C2 is 2000-2010 Years, C3 is 2010-2015, C4 is 2015-2020, D1 is less than 20 million yuan, D2 is 20-50 million yuan, D3 is more than 50 million. The corresponding address values are shown in Table3.

TABLE Ⅲ. WORK LABEL MATRIX

address value	***A1***	***A2***	***A3***	***A4***	***B1***	***B2***	***B3***	***B4***	***C1***	***C2***	***C3***	***C4***	***D1***	***D2***	***D3***
Work 1	0	1	0	0	0	1	1	0	0	0	0	1	0	0	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

In particular, when two (or more) tags are selected for a certain category of work, the category M+1 (...M+ n) is automatically created, and the category N (O, P...) will be set. The selected tags are stored one by one into their own categories and created new categories NN (O, P...).

e.g. The theme of a work is also ancient costume, love, and comedy. The category A: ancient costume, love, comedy, modern, and suspense are A1 A2 A3 A4 A5. Then created new categories K, L, K1, L1 which are ancient costume, K2, L2 which are love, K3, L3 which are comedy, K4, L4 which are modern, K5, L5 which are suspense. Then the address value of the work was 10000BC...M0100000100.

For works that have not been selected multiple times for this label, K and L were automatically saved as 0.

c) Database Q

The address value and the image array were stored in the database Q one by one. The image keyword matrix and the address value matrix have been formed into a matrix according to the correspondence relationship of each work, and have been stored in the database Q. Take " Wolf Warriors Ⅱ " as an example again, we knew that, its address value was “ 010001100001001”, and the image keyword a is patriotic, b is wonderful, c is stupid. Its value has been shown in the table below:

TABLE Ⅳ. DATABASE Q

	address value									image keyword				
	A1	*…An*	*B1*	*…Bn*	*C1*	*…Cn*	*…*	*M1*	*…Mn*	*a*	*b*	*c*	*…*	*n*
wk1	0	1,0,0,…	0	1,1,0,…	0	0,0,1,…	0,0,1,…	…	…	0.6508	0.16476	0.0986	…	…
wk2	…	…	…	…	…	…	…	…	…	…	…	…	…	…

d) Database S

The intimacy index of one-to-one mapping of attributes and features was calculated, and stored in the database S. As shown in the table below:

Intimacy index	*A1*	*…An*	*B1*	*…Bn*	*C1*	*…Cn*	*…*	*M1*	*…Mn*
a	$\alpha(A_1,a)$	…	$\alpha(B_1,a)$	…	$\alpha(C_1,a)$	…	…	$\alpha(M_1,a)$	…
b	$\alpha(A_1,b)$	…	$\alpha(B_1,b)$	…	$\alpha(C_1,b)$	…	…	$\alpha(M_1,b)$	…
c	$\alpha(A_1,c)$	…	$\alpha(B_1,c)$	…	$\alpha(C_1,c)$	…	…	$\alpha(M_1,c)$	…
d	$\alpha(A_1,d)$	…	$\alpha(B_1,d)$	…	$\alpha(C_1,d)$	…	…	$\alpha(M_1,d)$	…
e	$\alpha(A_1,e)$	…	$\alpha(B_1,e)$	…	$\alpha(C_1,e)$	…	…	$\alpha(M_1,e)$	…
f	……	…	……	…	……	…	…	……	…

TABLE Ⅴ. DATABASE S

2) "Feature - Image" model

The user selected the characteristics of the work to be analyzed, then the data in the database S from the background was called. The formula (6) or (7) would be chosen according to different situations. Consequently, the model calculated can calculate the comprehensive intimacy index between the work and different images. It selected the top five images that serve as predictive images.

3) "Attribute - Feature" model

The image attributes that you want to achieve in this work were selected, then the best label combination based on the intimacy index and dimensional influence was calculated.

e.g. A work wants to show the 'a' (kindness) image and the 'b' (cunning) image. Intimacy based on the multiplication of data in the database S and influence. The best intimacy in each feature tag could be found. The label, such as A1 (ancient costume) in A (type), B2 (Japan) in B (country). In order to obtain a picture showing the image of a (kindness) and the image of b (cunning), it is necessary to have the A1 (ancient costume) type, and B2 (Japan) to spread these two characteristics, that is, the best feature label combination of this work.

E. System Interfaces

1) Image analysis page

In addition to the basic information such as the type of film, the year of release, and the length of the film on the page, there is a very intuitive word cloud map to show some of the main communication images of film and television works abroad.

On this basis, there is an image analysis of film and television works. The analysis of the communication image was divided into two parts, the first part was the sentiment index of image, and the second part was the image proportion.

The user could click the gray button at the top to switch the propagation area, and click the data point or the fan-shaped area to display specific values.

Figure 2. Image analysis page.

2) Predicted results page

After selecting attributes such as the type, distribution area, release time and episode number in selecting properties page, you would get 7 prediction results, like "CG is bad", "good actor" etc. The value after the image is the comprehensive

intimacy index between the image and the attributes selected by the user.

Figure 3. Predicted results page.

3) Strategy suggestion page

In this interface, the user needed to select the image he wanted to spread, and the system would give corresponding strategy suggestions.

Figure 4. Strategy suggestion page.

F. Public approval

In order to test the reliability of the system, we conducted a public approval survey. There were 233 respondents, 75% of whom basically approved the system, 13% fully approved, and 12% disapproved.

IV. CONCLUSIONS

In this paper, an international image prediction system for film and TV works was established. This system analyzed the communication image by types, areas, episodes and so on, presented the analysis results with data visualization tools, and eventually made predictions for other film and TV series based on models established. Specifically, the system had the following characteristics: (1) It used sentiment analysis to calculate the sentiment index of each image and to give indicators for reference. (2) It established two prediction models, which respectively performed two functions of prediction image and strategy suggestion. The first prediction model-- the "feature-image" model-- was given the system's predicted communication image according to the attributes selected by the user. The user selected the characteristics of the work such as "ancient costume" in the type and "Japan" in the spreading area, then the system could predict "goodness", "gorgeous" and other images; the second prediction model-- "image-feature" model-- according to the expected image selected by the user, provided strategic recommendations. If the user wanted to get the image of "goodness", the system showed that the broadcast strategy was to set the type of work as an ancient costume play and the spreading area as Japan. The prediction system, based on objective data and with the model algorithm as the core, would make a more scientific prediction of the international image of film and television works, and played a reference role in the area of production and external communication of film and television works.

ACKNOWLEDGMENT

The work of this paper was supported by the National Natural Science Fund of China (Project 61901421), the National Social Science Fund of China (Project 19ZDA272), the Fundamental Research Funds for the Central Universities, State Key Laboratory of Media Convergence and Communication, Communication University of China.

REFERENCES

[1] Guan Wenhu. National Image Theory. Chengdu: Chengdu Science and Technology Humanities Publishing House, 2000. (In Chinese)

[2] Jiang Hongmin. From the perspective of cross-cultural communication, we explore the shaping of the national image of Chinese films going to the world. Shandong University, 2011. (In Chinese)

[3] Guo Qingguang. "Communication Tutorial", Renmin University of China Press, 1999 edition, p. 127 (In Chinese)

[4] Kanungo T, Mount D M, Netanyahu N S, et al.An efficient k-means clustering algorithm:analysis and implementation.IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24（7）:881-892

[5] Sun Jigui, Liu Jie, Zhao Lianyu.Study on Clustering Algorithms.Journal of Software,2008 (1):48-61. (In Chinese)

[6] WANG Binyu, LIU Wenfen, HU Xuexian, WEI Jianghong, Text clustering based on cosine distance selection initial cluster center. Computer Engineering and Applications, 2018, 54(10): 11-18. DOI:10.3778/j. Issn.1002-8331.1802-0108(In Chinese)

[7] Zhou Enbo, Mao Shanjun, Li Mei, et al. Improvement of GPU Acceleration PAM Clustering Algorithm Research and Application. Journal of Earth Information Science, 2017,19 (6):782-791. (In Chinese)

[8] HAN Bing, JIANG He, Improved K-medoids Algorithm Based on Similarity Calculation Formula. Computer and Modernization, 2019, (5): 113-117. DOI: 10.3969/j.issn.1006-2475.2019.05.021. (In Chinese)

[9] Shao Yan. Research on Chinese Family Ethics TV Drama. Shandong: Shandong Normal University, 2009. DOI:10.7666/d.Y1471234. (In Chinese)