Research on the Information Entropy Using Processing Square Matrix Method Based on Similarities Evaluation Model

Chupeng Cui*, Zhendong Jia, Ke Gu

College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012

*Corresponding author: chupengcui@jlu.edu.cn

Abstract. In the music history, there are plenty of splendid music artists. musicians in a certain genre may bring some varieties other genres, or influenced by them. this paper constructs a complex network model based on directional weighted graph. Specifically, we calculate the edge of the graph by referring the ideology of information entropy to describe the direct influence on two certain artists. Then, we take the 100 most influential artists and analyze their influence to get a direct influence square matrix. After that, we utilize the DEMATEL method to process this square matrix, and calculate the scale of the comprehensive influence square matrix about these musicians. This article uses Euclidean distance as the similarity metric, and measures the similarity between artists through 11 indicators. In order to facilitate the calculation, this paper uses the principal component analysis method to reduce the dimensionality of these indicators. In order to determine the influence relationship between genres, this paper uses the k-means algorithm to compare and analyze the clusters of artists with more than 10 followers.

1. Introduction

Music is a crucial part of human society. It is necessary that we need to hatch out a method measuring the evolution of music for the sake of digging out the development and effect of it in human society [1]. On the one hand, some objective factors, such as music genre, will have an impact on music creation, on the other, musicians will effect other. Therefore we aim to understand and assess how the music previously made influences new music and musicians by means of constructing related network.

2. Music influence network

We create a directional weighted influence network for the influence between two artists. In this network, each artist is a node and if one artists influence another, there is a weighted arc between them representing the influence value (a positive float number). By this network, we can not only discover who puts an impact

In the file influence_data, data is shown in a form of adjacency list. Considering the scale of the data, adjacency matrix is the best way to present the relationship. The adjacency matrix is a 2-dimension matrix utilized to analyze network relationships. Each entry of the matrix represents the connections between a couple of artists where one influence on another. In the adjacency matrix, it is easy to calculate
how many people does an artist influence on, or how many has influenced him. But simply link the point cannot represent the cohesion of their connection [2]. For example, one artist may admire another, so they share a close relationship. To simulate the stronger or weaker relationship, we must determine the criteria about edges in connections, that is, on what scale may one artist influence another. We define $m_{ij}$ as the impact one puts on another. Then, we construct a model to calculate this value.

2.1. Calculate the influence value

Considering two artists that has an influential connection. The influencer may contribute to a large amount of other artists, and the follower may be influenced by many masters. We consider 3 factors that contribute to this.

1. the influence between their genres,
2. the influence of their active years,
3. the year gap between two artists.

Suppose we now know that $i$ affects $j$. Next, we need to determine the magnitude of this influence. For $i$, his work may have an impact on all subsequent music. Assuming that $p_{1ij}$ represents the probability that $j$ may be affected by $i$'s influence on subsequent music, then the entropy of the information "i affects j" is $E_{1ij}$. Then, there is an equation

$$E_{1ij} = p_{1ij} \times log\left(\frac{1}{p_{1ij}}\right)$$

(1)

Obviously, we know that if the real impact is greater, the corresponding information entropy is smaller. Therefore, the reciprocal of information entropy can be used as an indicator of influence. Assuming that the unilateral influence that i may exert on j is $I_{ij}$, then there is an equation

$$I_{ij} = \frac{1}{E_{1ij}}$$

(2)

We use $I$ to describe the influence that influencer may put on it’s follower, and $A$ to describe the influence that follower may accept. We assume $i$ is the influencer and $j$ is his follower.

In order to calculate $p_{1ij}$, since we assume that the styles of composers in the same genre in the same time period are similar to the average style of this genre at that time, we use the probability of influence of the main genre of $j$ after the same genre in the period of $i$ as $p_{1ij}$. By counting all the impact relationship data [3], we can count the value of $p_{1ij}$.

Similarly, for $j$, he may be influenced by many people. Assuming that $p_{2ij}$ was brought by $i$ among the influences he received, then we can obtain $p_{2ij}$ by counting the number of influences of all the genres where $i$ belonged to the main genre of $j$ in the year of $j$. Similarly, there is an equation to calculate that $j$ may be affected by the era and genre of $i$

$$A_{ij} = \frac{1}{p_{2ij}} \times log\left(\frac{1}{p_{2ij}}\right)$$

(3)

As shown above, we have considered the first two impact factors. For the third impact factor, we take the empirical formula

$$\rho = log\left(1 + \frac{10}{\text{decade-gap}}\right)$$

(4)
As a prediction of year correlation. If two artists are in the same year, then decade_gap takes 1. Otherwise, decade_gap takes their year gap divided by 10. Finally, give the calculation formula of each edge weight

\[ m_{ij} = l_{ij} * A_{ij} * \rho \]  

(5)

In this way, we have a method for calculating the weight of each edge. But it can be seen that this link only considers the direct influence of two artists. There is no doubt that indirect effects exist. In order to be able to construct the comprehensive influence of the musician on his descendants, we try to construct a comprehensive influence matrix that can reflect the influence and changes the musician brings to all his colleagues.

2.2. Construct the comprehensive impact matrix

The sparse matrix with direct influence factors constructed by the above method reflects the direct influence of the artist. But the real influence of the artist may be profound and long-term. In order to calculate the comprehensive influence of an artist, therefore, we adopt the DEMATEL (Decision-making Trial and Evaluation Laboratory) method to obtain the comprehensive influence matrix of this matrix. The DEMATEL method is a method that uses graph theory and matrix tools to systematically analyze the comprehensive influence between nodes. After normalizing the adjacent square matrix, calculating all possible indirect influences of all relations and adding them up, the comprehensive influence of each element is obtained. By studying this comprehensive influence matrix, we can obtain the influentiality, acceptance, centrality and causality of each composer.

2.2.1. Calculate normalized direct influence matrix N

We selected the 100 most influential musicians, stripped out their influence relationship as a subgraph, and saved it in a new adjacency matrix. The first and second dimensions of this matrix correspond to each other and are the names of these 100 artists. Each element in the matrix represents the direct influence value of these two artists. If there is no direct influence, this value is 0. Therefore, this matrix is a 100*100 square matrix, which we call the direct influence square matrix M. First, we need to standardize the original relationship matrix to pay more attention to the strength of the impact. For the direct influence matrix M obtained before, we then use \( m_{ij} \) to represent the value within \( M(\text{ij}) \). Then, we calculate the maximum line in the formula

\[ \text{Maxvar} = \max(\sum_{j=1}^{n} m_{ij}) \]  

(6)

With the maximum value, we continue to calculate the specification normalized direct influence matrix N

\[ N = \left(\frac{m_{ij}}{\text{Maxvar}}\right)_{n \times n} \]  

(7)

2.2.2. Calculate the comprehensive relation matrix

Regarding the normalization that directly affects the square matrix N, it can be found that the k-th power of N \( N^k \) represents the indirect impact of any element in N through k passes. In order to calculate the comprehensive impact, we need to take all the indirect impacts into account, so there is a formula:

\[ T = (N + N^2 + N^3 + \cdots + N^k) = \sum_{k=1}^{\infty} N^k \]  

(8)

That is

\[ T = N(I - N)^{-1} \]  

(9)
Where $I$ is the identity matrix, and $(I - N)^{-1}$ is the inverse matrix of $(I - N)$. In this way, the comprehensive influence matrix $T$ can be obtained.

2.2.3. Calculate the metrics. **Influentiality, acceptance, centrality and causality** is the metrics of the influence of the four metrics in the system. We further calculate each element according to the value $t_{ij}$ of the comprehensive influence matrix $T$.

**Influentiality**

Influentiality measures the degree of comprehensive influence this artist has on other artists. Influentiality refers to the sum of the values of each row matrix of $T$, which represents the comprehensive influence value of the corresponding elements of each row on all other elements. This set is denoted as $D$, $D = (D_1, D_2, ..., D_n)$. Here is the equation

$$D_i = \sum_{j=1}^{n} t_{ij}, (i = 1, 2, 3, ..., n) \quad (10)$$

**Acceptance**

The greater the acceptance, the more the artist is affected. There is a formula: acceptance refers to the sum of the values of the columns of $T$, which means that the corresponding elements of each column are affected by the comprehensive influence of all other elements. This set is denoted as $C$, $C = (C_1, C_2, ..., C_n)$. The formula is

$$C_i = \sum_{j=1}^{n} t_{ji}, (i = 1, 2, 3, ..., n) \quad (11)$$

**Centrality**

Centrality represents the position of the factor in the evaluation index system and the magnitude of its effect. The greater the centrality, the greater its importance in the history of music. The centrality of the element $i$ is obtained by adding the Influentiality and acceptance of the element $i$ and denoted as $M_i$. So the formula is:

$$M_i = D_i + C_i \quad (12)$$

**Causality**

Causality is used to express the degree of originality an artist brings to the music industry. The larger the Causality, the more originality and freshness it brings to the music industry, and the more profound and lasting impact. We subtract an artist’s Influentiality and acceptance to get the causality of the element as $R_i$. That is:

$$R_i = D_i - C_i \quad (13)$$

2.3. Analysis

Through the above calculation, we can get the influence, acceptance, centrality and cause of each artist in this subgraph.
To illustrate this, we select some typical analysis. Among them, the green part of the pie chart is the degree of influence, and the gray part is the degree of influence. The size of the pie indicates the centrality of the composer. It can be concluded from the data that Bob Dylan is the most critical person in this subnet, and he deserves the Nobel Prize in Literature.

In this figure, the horizontal axis represents influentiality, the vertical axis represents acceptance, and the size of the circle represents causality. The artist at the bottom right of the line has a unique and novel
contribution to the history of music. In particular, the influence of Leonard Cohen and Bob Dylan is very profound. For the well-known, The Beatles and Neil Young, their influence is indelible. On the other hand, The Rolling Stones above the straight line have far-reaching influence, but compared to other artists on the list, they are more of bringing the rock style to more listeners rather than composers. Nirvana, on the other hand, seems to absorb and pass on the styles of these great artists on the list. Of course, since the artists we selected are the 100 most influential artists, it does not mean that their real influence is very small. On the contrary, they still play a vital role in music history and their works are unforgettable.

3. Music Similarity Evaluation Model

3.1. Use principal component analysis to obtain evaluation indicators

By analyzing the data set, we found 11 indicators can be used to evaluate the artist, which are: acousticness, instrumentalness, liveness, speechiness, duration_ms, popularity, danceability, energy, valence, tempo and loudness. In order to optimize the comprehensive simplification of the multivariate cross-sectional data table and further reduce the amount of calculation, we use **principal component analysis** to reduce the main indicators of the song.

**Standardize data**

Suppose there are m variables for principal component analysis and n evaluation objects, the value of the j(j<=m) index of the i(i<=n) evaluation object is \( x_{ij} \). We first convert each index value \( x_{ij} \) into a standardized index \( \bar{x}_{ij} \)

\[
\bar{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}
\]  

\[
\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}, S_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}
\]

**Calculate the correlation coefficient matrix R**

\[
R = \begin{pmatrix} r_{ij} \end{pmatrix}_{m \times m}
\]  

\[
r_{ij} = \frac{\sum_{k=1}^{n} x_{kj} \bar{x}_{ki}}{n-1}
\]

**Calculate eigenvalues and eigenvectors**

To calculate a correlation coefficient matrix \( R \) eigenvalues \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0 \) and corresponding eigenvectors \( \mu_1, \mu_2, \ldots, \mu_m \), eigenvectors obtained consisting of \( m \) new indicator variables.

\[
\begin{cases}
  y_1 = \mu_{11} \bar{x}_1 + \mu_{21} \bar{x}_2 + \cdots + \mu_{n1} \bar{x}_n \\
  y_2 = \mu_{12} \bar{x}_1 + \mu_{22} \bar{x}_2 + \cdots + \mu_{n2} \bar{x}_n \\
  \vdots \\
  y_m = \mu_{1m} \bar{x}_1 + \mu_{2m} \bar{x}_2 + \cdots + \mu_{nm} \bar{x}_n
\end{cases}
\]
3.2. Calculate the contribution value and select the principal component

Next, we calculated the information contribution rate $b_j$ of the principal component $y_j$ and the cumulative contribution $\alpha_p$ of the principal components $y_1, y_2, ..., y_p$, and selected the indicator variable group with a cumulative contribution of more than 80% as the new indicator.

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{m} \lambda_k} (j = 1, 2, ..., m)$$

$$\alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{m} \lambda_k}$$

After calculation, we find that $\alpha_p > 0.8$ when $p=6$, so we choose the first 6 index variables $y_1, y_2, ..., y_6$ to replace the original 11 index variables as the new index to evaluate the characteristics of the artist.

3.3. Euclidean distance for similarity measurement

The closer the Eigenvalues between different artists are, the greater the similarity between artists. Therefore, after obtaining the six new indicators for evaluating artist characteristics [4], we map an artist to a point in a six-dimensional space, and his six feature values correspond to the coordination in the space. After that, we can use the Euclidean distance between two points to measure the similarity between two artists. The metric calculation formula is as follows:

$$\text{dist}(X, Y) = \sqrt{\sum_{i=1}^{6} (x_i - y_i)^2}$$

3.4. Cluster analysis using k-means algorithm

After obtaining the Euclidean distance between all points, we use the k-means algorithm to divide the singers into different groups according to their music similarity, then divide the singers with higher music similarity into the same group. Specific steps are as follows:

3.4.1. Selection of K value. The K value represents the number of samples in a sample group, and the optimal solution of the k value is to minimize the cost function, so we use the elbow method to determine the optimal k value for clustering. First, we use the SSE(sum of the squared errors) as an indicator, then the calculation formula is as follows:

$$\text{SSE} = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

And get the graph of the relationship between SSE and k value.
Figure 3. The relationship between the SSE and the k value

As the number of clusters k increases, the sample division will be more refined, the degree of aggregation of each cluster will gradually increase, so the SSE will gradually decrease. When k is less than the true number of clusters, since the increase of k will greatly increase the degree of aggregation of each cluster, the SSE will decrease greatly. But when k reaches the number of true clusters, the return on the degree of aggregation obtained by k will reduce rapidly. Consequently, the decline of SSE will decrease sharply, and then SSE will level off as the value of k continues to increase. Therefore, the k value we determine is 7.

3.4.2. Select initial centroid. We first randomly select k samples from the data set $D = \{x_1, x_2, \ldots, x_m\}$ as the initial k centroid vectors $\{\mu_1, \mu_2, \ldots, \mu_k\}$. For $n=1,2,\ldots,N$, we first initialize the cluster partition $C$ as

$$C_t = \emptyset$$

(23)

Use the calculated similarity metric to get the distance between the sample $x_i$ and each centroid vector $\mu_j (j = 1,2,\ldots,k)$. Mark the smallest $x_i$ as the category $\tilde{\lambda}$ corresponding to $d_{ij}$, and then update $C_{\tilde{\lambda}} = C_{\tilde{\lambda}} \cup \{x_i\}$. Then recalculate new centroids for all sample points in $C_j$

$$\mu_j = \frac{1}{|C_j|} \sum_{x \in C_j} x$$

(24)

3.4.3. Obtain clustering results after iterative calculation. We iteratively calculate the centroids of all sample points, until all k centroid vectors never change further. In this way, we get the 7 cluster divisions $C = \{C_1, C_2, \ldots, C_7\}$ after clustering, and each cluster division represents similarity artist collection. There are two clusters with only one artist. We treat these two clusters as noise without consideration, so the actual clusters examined are five.
3.5. **Artist similarity evaluation standard**

According to the music similarity measurement evaluation model, we group the artists with high similarity into the same group based on the similarity measurement between artists, which we call similar groups. Therefore, we got the following similarity evaluation criteria:

For artists of the same genre, if most of them are concentrated in one group, the similarity of the artists of the genre is high, and if they are scattered in different groups, the similarity of the artists of the genre is low. For different genres, as far as artists are concerned, if they are concentrated in the same similar group, it means that the similarity of artists between these genres is relatively high, and if the artists of different genres are concentrated in different similar groups, it means that the similarity of artists between these genres is relatively low.

4. **Conclusion**

We counted the genres of the artists in 5 similar groups and the proportion of the number of people of that genre in a similar group. We can get that the genres with higher similarity of artists in the genre are Classical, Easy Listening, New Age, Stage & Screen and Reggae, and the genres with lower similarity of artists in the genre are Electronic, International, Pop/Rock; Among different genres, the similarity of artists between the genres Classical, Easy Listening, New Age and Stage & Screen is relatively high. The artists of Electronic, Reggae, and R&B are also relatively similar, and most of the genres are similar. The degree is low.

**References**

[1] Tseng, M. L. (2009). Application of anp and dematel to evaluate the decision-making of municipal solid waste management in metro manila. Environmental Monitoring & Assessment, 156(1-4), 181-197.

[2] Liang, J., Shi, Z., D., L. I., & Wierman, M. J. (2006). Information entropy, rough entropy and knowledge granulation in incomplete information systems. International Journal of General Systems, 35(6), 641-654.

[3] Wong, J. A. H. A. (1979). Algorithm as 136: a k-means clustering algorithm. Journal of the Royal Statistical Society, 28(1), 100-108.

[4] Wagstaff, K. (2001). Constrained k-means clustering with background knowledge. Proceedings of ICML-2001.