Research Article

Design of a Music Recommendation Model on the Basis of Multilayer Attention Representation

Wei Lu

Conservatory of Music of Hefei Normal University, Hefei 230601, China

Correspondence should be addressed to Wei Lu; luwei@hfnu.edu.cn

Received 14 October 2021; Revised 22 November 2021; Accepted 29 November 2021; Published 21 January 2022

Academic Editor: Rahman Ali

Copyright © 2022 Wei Lu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Current music recommendation systems can explore the general relationship between the users and songs to recommend music to the users; however, they cannot distinguish the different preferences of different users for the same song. For example, a user may like a song because of the singer, while another user will like it not for the singer but just because of the composition of the song or its melody. A recommender system that knows this difference would be more effective in recommending music to the users. To this end, this paper proposes a music recommendation model based on multilayer attention representation, which learns song representations from multidimensions using user-attribute information and song content information, and mines the preference relationship between users and songs. In order to distinguish the differences in user preferences for multidomain features of songs, a feature-dependent attention network is designed; in order to distinguish the differences in user preferences for different historical behaviors and to explore the temporal dependence of user behaviors, a song-dependent attention network is designed. Finally, the SoftMax function is used to calculate the distribution of users' preferences for candidate songs and is used to generate recommendations. The experimental results on 30Music and MIGU datasets show that the proposed model achieves significant improvement in recall and MRR compared with the current recommendation models.

1. Introduction

The popularity of streaming music has brought users great freedom and unlimited access to music content, such as Last.fm, Pandora, QQ Music, NetEase Cloud Music, and MIGU Music. For these music platforms, one of the main challenges is how to recommend songs to target users, let users quickly find their favorite songs from the massive library, and improve their listening experience. In recent years, with the collection of massive user-listening behavior data, recommendation systems have been successfully applied in social multimedia platforms. Therefore, streaming music platforms have also introduced recommendation systems to varying degrees, mining users' music preferences based on their listening behavior and recommending songs that users may like. Collaborative filtering based on nearest neighbors is the first recommendation algorithm used in song recommendation, which generates recommendations based on calculating the nearest neighbors of users or songs in terms of

behavior and achieves good recommendation results. However, user-behavior data in streaming music platforms usually obey the long-tail distribution, with 20% of popular songs aggregating 80% of listening behaviors [1], which leads the platform to recommend a large number of disliked popular songs to users in a specific scenario, i.e., the popularity bias, while those songs that are rarely listened to by users or new on the shelves are difficult to be recommended to target users. In addition, the number of songs that users have listened to is extremely small compared to the massive music library, which tends to cause the problem of the sparsity of user behavior data. To solve these problems, researchers proposed a model-based collaborative filtering method [2, 3], which uses a vector of intermediate implicit factors to connect users and songs and predicts users' preferences for songs by dot product operations. Although this approach reduces the data sparsity problem and popularity bias to some extent, it is still based on a single interaction behavior of users and ignores the difference in users' preferences for song content features.

The streaming music platform contains not only a large amount of user behavior data but also records metadata of songs, such as artists, lyrics, genres, covers, and audio. These metadata can be used to design new recommendation models and improve the recommendation accuracy of the models. The content filtering method [4] makes up for the shortage of collaborative filtering by mining the content features of songs that users have listened to recommend songs that are similar to them. However, content filtering ignores the “synergy law” among users because similar songs have similar properties and similar users have similar preferences. Therefore, how to combine the advantages of both approaches to construct more efficient recommendation algorithms has become an urgent problem for academia and industry. The early brute-force fusion model is a greedy strategy to select the songs with the highest expected user ratings to generate a safer recommendation list, which brings long-term suboptimal problems because users’ music preferences are estimated based on existing knowledge, ignoring the influence of those unknown knowledge on users’ preferences, such as temporal dependencies and contextual information.

In recent years, domestic and foreign researchers have also proposed hybrid recommendation algorithms based on session and context [5, 6] to fully exploit the influence of unknown knowledge on music preferences and improve recommendation quality. These methods only consider the user’s contextual information and temporal relationships in the current session at the global level and lack the understanding of fine-grained features. For example, users have different interest points for different songs in the list of listened songs, and there are differences in what different users pay attention to the same song.

To this end, this paper proposes a recommendation model (HARM) based on multilayer attention representation, which uses user-attribute information and song content information to learn the embedded representation of songs from multiple dimensions and mine the user’s preference features for songs. Specifically, in order to distinguish the differential preferences of users on multidomain features of songs, a user-feature-dependent attention network is designed. To distinguish the differences of different historical behaviors on user preferences and to explore the temporal dependencies of behaviors, a song-dependent attention network is designed. Finally, the SoftMax function is used to calculate the distribution of users’ preferences for candidate songs and is used to generate recommendations. In addition, to be able to extract audio content features, CNN networks are pretrained using genre label data of songs and then combined with recommendation tasks for supervised learning optimization for learning song audio representations.

The main contributions of the paper include the following:

We propose a song recommendation model based on multilayer attention representation which is comprised of three parts.

- (1) First, we build a user-based attention network to learn song representations.
- (2) Then, we build a song-based attention network to learn user preference representations based on the song representations listened to by users.
- (3) Finally, we use the learned user-preference representations and song representations to predict the probability of users’ preferences for candidate songs and recommend the top N songs with higher scores to users.

The rest of the paper is organized as follows:

Section 2 discusses the relevant work in the subject of recommender systems for music and other purposes. Section 3 discusses in detail the proposed recommendation model based on multilayer attention representation. Section 4 describes the experimental setup and the experiments carried out. Section 5 analyzes the experimental results, and section 6 is the conclusion of our work.

2. Related Work

Both collaborative and content filtering-based models rely on behavioral data of user-item interactions, which leads these models to focus only on the long-term static preferences of users. There are extremely strong correlations and causality in the user’s behavioral sequences. Sequence mining is the earliest approach to mining temporal relationships, inferring the future behavior of users by mining the sequential relationships between items in a session. The Markov chain model is the most classical method of sequential relationship mining [7], which predicts the next behavior of a household by calculating the probability of single-step transfer between behaviors. User behavior is often influenced by multistep dependencies, and in order to mine such influences, higher-order Markov chain models [8] consider more dependencies of previous behaviors. However, the large number of states in the multistep dependency calculation process leads to the excessive computational effort. In addition, the user-behavior data is extremely sparse, leading to inaccurate recommendation results. To solve the computational complexity problem under multiple behaviors, Li et al. [9] proposed a state-space model based on the sequence of item attributes to alleviate the problem of state-space explosion; Kolivand et al. [10] proposed a temporal recommendation model based on matrix decomposition to generate the next recommendation item based on the interaction between session and candidate items; Gao et al. [11] proposed a FISM model that performs matrix decomposition of the item-item colinear matrix without learning explicit user representations. In general, these Markov models based on matrix decomposition only consider the action relations in lower-order cases and cannot solve the interactions in higher-order cases, and at the same time, they ignore the behavioral order dependencies within and between sessions. Recommendation systems have widely been used in different areas for different services and

applications, such as wellness [12] and recommendations of machine learning algorithms for learning model creation [13]. They have also been widely used for video recommendation on online platforms. Davidson et al. [14] discussed the video recommendation system used by YouTube to recommend videos to users. Deldjoo et al. [15] focused on the visual features of videos in the context of recommendation systems and proposed a content-based video recommendation system. Sun [16] proposed VideoReach, a video recommendation system by finding a list of relevant videos in terms of relevance such as textual, visual and aural relevance, and user click through.

3. Recommendation Model Based on Multilayer Attention Representation

In this paper, we improve the quality of song recommendation by building a recommendation model based on multilayer attention representation. The proposed model is comprised of three parts including song coding, user coding, and prediction. First, we build a user-based attention network to learn song representations; then, we build a song-based attention network to learn user-preference representations based on the song representations listened to by users; finally, we use the learned user-preference representations and song representations to predict the probability of users' preferences for candidate songs and recommend the top N songs with higher scores to users. The framework of the harm model is shown in Figure 1.

3.1. Song Coding. At present, the coding method of songs usually uses embedded representation technology to map the unique representation of users listening to songs into dense semantic feature vectors and take them as the coding features of songs. This coarse-grained coding method cannot describe the fine-grained features of songs. This leads to inaccurate song recommendations. Users' preferences for a song are usually affected by multidomain features, such as song genre, singer, and audio melody. In addition, different users tend to pay different attention to the preferences of the same song. Therefore, it is necessary to distinguish users' different preferences for song features from the fine-grained level. Therefore, the song coding in the harm model consists of three parts: embedded representation of discrete features, extraction of audio semantic features, and user-based attention network.

For the discrete eigenvalue x_{ij} of the j -th domain of song I , firstly, the unique heat representation is used to generate a high-dimensional sparse eigenvector $f_{ij} \in \mathbb{R}^{1 \times |f_{ij}|}$.

$|f_{ij}|$ represents the number of discrete values in the j -th domain. Using the embedded representation technology, the sparse features are mapped to f_{ij} to the dense semantic space, that is, $\mathbf{e}_{ij} = \mathbf{f}_{ij} \mathbf{E}_j$, $\mathbf{E}_j \in \mathbb{R}^{|f_{ij}| \times N_j}$ is the embedded representation matrix of the j -th domain, which is obtained through model training and optimization, and N_j represents the embedded dimension. For the audio content, it is represented by x_{ik+1} . Firstly, the audio signal is transformed into

the frequency domain by discrete Fourier transform. The corresponding color sound spectrum is generated for each song. The depth of the sound spectrum represents the sound size at this frequency, and the length and width represent the time and frequency. Due to different playing time of each song, the color sound spectrum of each song is divided into multiple $256 \times \text{two hundred and fifty-six}$. Finally, the semantic content representation is extracted from the song audio by using the pretrained VGG16 network model based on ImageNet dataset e_{ik+1} .

In the actual scene, different users pay attention to the same song from different perspectives. For example, some users pay attention to the singer of the song, while others pay attention to the style or melody of the song. In addition, even the same characteristics of the same song contribute significantly to different user preferences. Therefore, in order to distinguish the importance of the characteristics of different domains of songs to different users, this paper proposes a user-based attention network based on the user's feature representation. The contribution of different domain features is recognized and strengthened. The user's feature representation is composed of multiple user features. Therefore, firstly, the user's ID, age, gender, and other features are mapped to the dense representation space and then spliced to obtain the user-embedded feature vector. $e_{it} \in \mathbb{R}^{1 \times D_e}$, where D_e represents the dimension of the feature vector. Then, the user-preference vector is calculated by using the full connection layer. $q \in \mathbb{R}^{1 \times D_q}$, where D_q represents the dimension of the user-preference vector. The calculation formula is

$$\mathbf{q} = \text{ReLU}(\mathbf{V} \times \mathbf{e}_u + \mathbf{b}_u), \quad (1)$$

where $b \in \mathbb{R}^{1 \times D_q}$, $v \in \mathbb{R}^{D_e \times D_q}$ is the model parameter. In this network, the user-preference vector q is used as the query, the embedded representation of each domain in the song is used as the key-value pair, and the attention mechanism is used to calculate the weight of each domain feature. Specifically, the formal calculation method of the weight a_j of the j -th domain of song I is

$$\mathbf{r}_i = \sum_{k=1}^{K+1} a_j \mathbf{e}_{ij}. \quad (2)$$

All songs in the user-listening sequence are encoded according to the above method so as to obtain the song feature representation sequence of the user-listening sequence, $R = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_L]$, where L represents the sequence length, $\mathbf{R} \in \mathbb{R}^{N_j \times L}$.

3.2. User Code. The purpose of user coding is to learn the user's preference representation from the music feature representation listened to by the user. Many methods can be used to learn the user's preference representation, including the method based on weighted pooling [17], the method based on the attention mechanism [18], and the method based on the cyclic neural network [19]. In this paper, the user-based attention network is used to learn the user-preference representation. Specifically, the self-attention

FIGURE 1: Framework of harm model.

network is used to learn the fine-grained timing dependence between users listening to songs.

In order to capture the timing dependence of the user's listening behavior, each position L in the user's listening sequence is encoded, represented by c_l , and all position codes of the listening sequence are represented by matrix $c = [c_1, c_2, \dots, c_L]$, $C \in N_j \times L$. Therefore, the song representation with position coding in the user's listening sequence can be represented as $F = C \oplus R$, where $\oplus$ is an element level addition operation, $F \in N_j \times L$.

Considering the song feature representation f in the sequence as the query Q , key K and value of the self-attention mechanism, and using the fully connected network to transform f nonlinearly, the relationship weight matrix between music listening behavior is expressed as

$$\begin{aligned} A &= \text{Softmax}\left(\frac{QK^T}{\sqrt{N_j}}\right), \\ Q &= \eta(F^T W^Q + B_q), \\ K &= \eta(F^T W^K + B_k). \end{aligned} \quad (3)$$

Among them, $\eta(\cdot)$ is a nonlinear activation function. In this paper, the ReLU function is used to increase the

nonlinear ability; $W^Q \in N_j \times N_i$, $W^K \in N_j \times N_i$, $B_q \in L \times N_i$, $B_k \in L \times N_i$ is the model parameter; $\sqrt{N_j}$ scale is the dot multiplication operation. The output of equation (4) is $L \times N_i$. On this basis, the output of the attention network is a matrix of $L \times N_i$, which is defined as

$$O = AF^T. \quad (4)$$

In order to obtain the user's global preference representation, this paper splices the behavior representation of equation (5) at each time and then uses the fully connected network to learn the user's global preference representation, which is expressed as

$$u_i = \text{concat}(O_1, O_2, \dots, O_L)W^u, \quad (5)$$

where concat is a vector splicing function, $W^u \in L \times N_i$.

3.3. Prediction and Training. For the target song S_t in the candidate set, firstly, the target song t is encoded by using the method in Section 3.1 to obtain the encoded representation r_t of the song. In this paper, the user's preference probability for the target song is calculated by using the SoftMax function, which is expressed as

$$P(S_t | \mathbf{u}_i) = \frac{\exp(\mathbf{u}_i \mathbf{r}_t^T)}{\sum_l^{[V]} \exp(\mathbf{u}_i \mathbf{r}_l^T)} \quad (6)$$

where V is the number of songs in the candidate set, and $\mathbf{u}_i$ is the user-preference representation vector. Given the training set $\chi = \{\langle H_i^L, \mathbf{u}_i, S_i \rangle\}$, the joint probability distribution can be expressed as

$$P_\theta(\chi) = \prod_{x \in \chi} \frac{\exp(\mathbf{u}_i \mathbf{r}_t^T)}{\sum_l^{[V]} \exp(\mathbf{u}_i \mathbf{r}_l^T)} \quad (7)$$

Parameter $\theta = \{\mathbf{E}, \mathbf{V}, \mathbf{W}_q, \mathbf{W}^e, \mathbf{W}^K, \mathbf{W}^U, \mathbf{b}_q, \mathbf{B}_q, \mathbf{B}_k\}$ can be learned by minimizing the conditional log likelihood probability, i.e.,

$$\mathcal{L} = \sum_{x \in \chi} P(S_t | \mathbf{u}_i) - \lambda \theta. \quad (8)$$

Directly optimizing equation (8) is very time consuming because it is necessary to calculate the SoftMax function for all songs. Therefore, referring to work [20], this paper uses the negative sampling technology to improve the model optimization speed and approximate the original objective function L . The target loss function after transformation is

$$\mathcal{L} = \sum_{x \in \chi} \log \sigma(\mathbf{u}_i \mathbf{r}_t^T) + k \cdot E_{j' \sim G_j} [\log \sigma(\mathbf{u}_i \mathbf{r}_{j'}^T)] - \lambda \theta. \quad (9)$$

Among them, σ is the activation function, K is the number of negative samples sampled according to the noise distribution G_j , and G_j is expressed by the distribution frequency of users listening to songs. Finally, the random gradient descent method is used to train the model.

4. Experiments

4.1. Experimental Data. This paper selects two datasets, 30music and MIGU, to verify the proposed model. 30music dataset [1] is an extended dataset based on last.fm, which contains attribute information of users and songs, as well as complete session information. The MIGU dataset is a private dataset provided by MIGU music. It contains the song metadata and user-characteristic data required by the model. In order to facilitate the experiment, the original dataset is cleaned. For the 30music dataset, the playback events have been deleted that lack the user or song attribute information in the original session information. Only the session information with playback events greater than 10 is retained. Then, the audio file corresponding to the song is captured on last.fm. Finally, 313,941 sessions and 4,267,295 playback records are generated, including 13,535 users and 27,302 songs, with an average of 13 playback events per session. For the MIGU dataset, only the complete records of user or song attribute information in playback events are retained. It is ensured that the time interval between adjacent playback events in the session is less than 200 s and retained the playback events with a listening time less than 30% or 20 s. Finally, 1,374,236 sessions and 17,621,958 playback records are generated, including 41,407 users and 1,926,590 songs, with an average of 12 playback events per session.

4.2. Comparison Method. This paper selects the following advanced recommendation algorithms to compare with the proposed model.

- (1) NARM [17]: It is a hybrid encoder, which uses the attention mechanism to model the user's continuous behavior, capture the user's intention from the current session, and then combine the intention into a unified session representation to calculate the recommendation score of each candidate.
- (2) SASRec [21]: A sequence model based on self-attention. The model first identifies relevant items in the user's operation history and then uses them to predict the next preferred item.
- (3) Gru4rec [22]: RNN is applied to the field of a recommendation system for the first time. The model introduces session parallelism, output sampling based on small batch, and sorting the loss function.
- (4) Improved-gru4rec [23]: Based on the gru4rec model, strategies such as data expansion and considering the time change of user behavior are added to improve the model performance.

4.3. Evaluation Method. In this paper, recall and MRR indicators are used to measure the recommendation effect of the proposed model and the comparison model. Recall: it is defined as the proportion between the songs in the recommendation list and all the songs users like.

$$\text{Recall@N} = \frac{1}{|D^{\text{Test}}|} \sum_{u \in D^D} \frac{|S_u^{\text{pre}} \cap D_u^{\text{Test}}|}{|D_u^{\text{Test}}|}. \quad (10)$$

Among them, S_u^{pre} represents the list of the first n songs recommended by user u , D_u^{Test} represents the list of songs in the test set, and $|D_u^{\text{Test}}|$ represents the number of samples of user u in the test set.

MRR: recall only measures whether the user's favorite songs appear in the recommendation list without considering the ranking factor. MRR is a sort measurement index, which is defined as

$$\text{MRR} = \frac{1}{|U|} \sum_i^{|U|} \frac{1}{\text{rank}_i}. \quad (11)$$

Among them, $|U|$ represents the number of users, and rank_i represents the ranking position of the i th song in the recommendation list of user u in the real list.

The model training is carried out on a 4-core high-performance NVIDIA GPU server, and each GPU contains 24 GB video memory. The comparison method and the model proposed in this paper are implemented based on the TensorFlow framework of python, and the model super-parameters are obtained by the cross-validation method.

5. Experimental Results and Analysis

5.1. Comparative Analysis. In order to evaluate the accuracy of the harm model and comparison model, tests were carried out on 30music and MIGU datasets, respectively. Recall and

FIGURE 2: Recall results of the harm model and comparison model on 30music and MIGU datasets. (a) 30music dataset. (b) MIGU dataset.

FIGURE 3: MRR results of the harm model and comparison model on 30music and MIGU datasets. (a) 30music dataset. (b) MIGU dataset.

MRR were used to measure the recommended quality. The number of recommended songs n ranged from 5 to 30 with an interval of 5. In order to eliminate the instability of single experimental results, this paper repeated 10 experiments and took the average of experimental results as the final evaluation result.

Figure 2 shows the recall experimental results of the harm model and comparison model on 30music and MIGU datasets. The experimental results show that recall @ 20 improves 49% on the 30music dataset and 44% on the

MIGU dataset compared with the traditional model (BPR). Compared with the time series models (gru4rec and improved-gru4rec), recall @ 20 improves 44% and 32% on the 30music dataset and 33% and 24% on the MIGU dataset, respectively. Compared with the attention model (NARM and Sasrec), recall @ 20 increased by 28% and 21% on the 30music dataset and 24% and 17% on the MIGU dataset, respectively.

The MRR indexes of the models are also evaluated. The performance of MRR indexes of each model on 30music and

MIGU datasets is shown in Figure 3. The MRR values of the harm model are better than the comparison method. Compared with SASRec, they are increased by 10% and 5.6%, respectively on 30music and MIGU datasets. This is because, in music scenes, user preferences are often affected by multidomain characteristics. Different users also pay different attention to the characteristics of the same song. The harm model proposed in this paper can more finely distinguish the preference differences of different users for the same song and finely describe the song feature representation so as to improve the quality of song recommendation.

In addition, from the experimental results of the two datasets, the index values of the harm model in the 30music dataset are higher than the MIGU dataset because the MIGU dataset is collected from the real production environment, and the data sparsity is higher, while the 30music dataset is obtained after specific cleaning.

6. Conclusions

This paper proposes a music recommendation model based on multilayer attention representation. The existing music recommendation systems cannot distinguish between the nature of users' preferences such as someone's choice of the singer or the composition of the song to effectively recommend music. Our proposed song recommendation model is based on multilayer attention representation in which we build a user-based attention network to learn song representations. Then, we build a song-based attention network to learn user-preference representations based on the song representations listened to by users, and finally, we use the learned user-preference representations and song representations to predict the probability of users' preferences for candidate songs and recommend the top N songs with higher scores to the user. The experimental results on 30music and MIGU datasets show that the experimental results of the proposed model perform better in terms of recall and MRR than the competitor methods. On 30music and MIGU datasets, the recall index of the harm model is improved by 49% and 44%, respectively compared with BPR; the MRR index of the model is 10% and 5.6% higher than the second-ranked SASRec, respectively.

Data Availability

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

Conflicts of Interest

The author declares no conflicts of interest.

References

- [1] S. Shang, J. Liu, and Y. Yang, "Multi-layer transformer aggregation encoder for answer generation," *IEEE Access*, vol. 8, Article ID 90410, 2020.
- [2] A. Macanovic, C. Marquette, C. Polychronakos, and F. Lawrence, "Impedance-based detection of DNA sequences using a silicon transducer with PNA as the probe layer," *Nucleic Acids Research*, vol. 32, no. 2, 2004.
- [3] X. Xia and J. Yan, "Construction of Music Teaching Evaluation Model Based on Weighted Naïve Bayes," *Scientific Programming*, vol. 2021, Article ID 7196197, 9 pages, 2021.
- [4] G. Cai, G. Lyu, Y. Lin, and Y. Wen, "Multi-level deep correlative networks for multi-modal sentiment analysis," *Chinese Journal of Electronics*, vol. 29, no. 6, pp. 1025–1038, 2020.
- [5] L. Yang and H. Hu, "Adaptive syncretic attention for constrained image captioning," *Neural Processing Letters*, vol. 50, no. 1, pp. 549–564, 2019.
- [6] Y. Geng, Y. Lian, M. Zhou, Y. Kong, and Y. Zhu, "Exploiting multigranular salient features with hierarchical multi-mode attention network for pedestrian re-identification," *Journal of Visual Communication and Image Representation*, vol. 73, Article ID 102914, 2020.
- [7] M. Guimberteau, A. Ducharne, P. Ciais, and J. P. Boisier, "Testing conceptual and physically based soil hydrology schemes against observations for the Amazon Basin," *Geoscientific Model Development*, vol. 7, pp. 1115–1136, 2014.
- [8] Y. Xiao, "Research on visual image texture rendering for artistic aided design," *Scientific Programming*, vol. 2021, Article ID 1190912, 8 pages, 2021.
- [9] F. Li, Q. Dai, W. Xu, and G. Er, "Multilabel neighborhood propagation for region-based image retrieval," *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1592–1604, 2008.
- [10] H. Kolivand, Z. Noh, and M. S. Sunar, "A quadratic spline approximation using detail multi-layer for soft shadow generation in augmented reality," *Multimedia Tools and Applications*, vol. 73, no. 3, pp. 1225–1245, 2013.
- [11] G. Gao, Y. Yu, J. Xie, J. Yangd, M. Yange, and J. Zhang, "Constructing multilayer locality-constrained matrix regression framework for noise robust face super-resolution," *Pattern Recognition*, vol. 110, Article ID 107539, 2020.
- [12] M. Afzal, S. I. Ali, R. Ali et al., "Personalization of wellness recommendations using contextual interpretation," *Expert Systems with Applications*, vol. 96, pp. 506–521, 2018.
- [13] R. Ali, S. Lee, and T. C. Chung, "Accurate multi-criteria decision making methodology for recommending machine learning algorithm," *Expert Systems with Applications*, vol. 71, pp. 257–278, 2017.
- [14] J. Davidson, B. Liebald, J. Liu et al., "The YouTube video recommendation system," in *Proceedings of the Fourth ACM Conference on Recommender Systems*, pp. 293–296, Barcelona, Spain, 2010 September.
- [15] Y. Deldjoo, M. Elahi, P. Cremonesi, F. Garzotto, P. Piazzolla, and M. Quadana, "Content-based video recommendation system based on stylistic visual features," *Journal on Data Semantics*, vol. 5, no. 2, pp. 99–113, 2016.
- [16] S. Sun, "Evaluation of potential correlation of piano teaching using edge-enabled data and machine learning," *Mobile Information Systems*, vol. 2021, Article ID 6616284, 11 pages, 2021.
- [17] Á. García-Crespo, R. Colomo-Palacios, J. M. Gómez-Berbís, and B. Ruiz-Mezcua, "SEMO: a framework for customer social networks analysis based on semantics," *Journal of Information Technology*, vol. 25, no. 2, pp. 178–188, 2010.
- [18] A. Da'U, N. Salim, and R. Idris, "Multi-level attentive deep user-item representation learning for recommendation system," *Neurocomputing*, vol. 433, pp. 119–130, 2021.
- [19] X. Y. Gao, Y. Z. Wang, C. X. Zhang, and J. Q. Lu, "Multi-head self-attention for 3D point Cloud classification," *IEEE Access*, vol. 9, no. 99, Article ID 18137, 2021.

- [20] H. Ohkushi, T. Ogawa, and M. Haseyama, “Music recommendation based on relationship between human motions and music pieces: a note on its automation and accuracy improvement,” *Ite Technical Report*, vol. 35, no. 9, pp. 83–87, 2011.
- [21] A. Szmul, T. Matin, F. V. Gleeson, J. A. Schnabel, V. Grau, and B. W. Papież, “Patch-based lung ventilation estimation using multi-layer supervoxels,” *Computerized Medical Imaging and Graphics*, vol. 74, pp. 49–60, 2019.
- [22] D. Jannach and M. Ludewig, “When recurrent neural networks meet the neighborhood for session-based recommendation,” in *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pp. 306–310, Como, Italy, August 2017.
- [23] P. Ren, Z. Chen, J. Li, Z. Ren, J. Ma, and M. De Rijke, “RepeatNet: a repeat aware neural recommendation machine for session-based recommendation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 4806–4813, Honolulu, HI, USA, February 2019.