IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 20, 2020, accepted February 16, 2020, date of publication March 4, 2020, date of current version March 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2978281

Context-Aware Network Analysis of Music Streaming Services for Popularity Estimation of Artists

YUI MATSUMOTO[1], (Student Member, IEEE), RYOSUKE HARAKAWA[2], (Member, IEEE), TAKAHIRO OGAWA[3], (Senior Member, IEEE), AND MIKI HASEYAMA[3], (Senior Member, IEEE)

[1]Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan
[2]Department of Electrical, Electronics, and Information Engineering, Nagaoka University of Technology, Nagaoka 940-2188, Japan
[3]Faculty of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan

Corresponding author: Yui Matsumoto (matsumoto@lmd.ist.hokudai.ac.jp)

This work was supported in part by the MIC/SCOPE #181601001 and in part by the Photo-Excitonix Project of Hokkaido University.

ABSTRACT A novel trial for estimating popularity of artists in music streaming services (MSS) is presented in this paper. The main contribution of this paper is to improve extensibility for using multi-modal features to accurately analyze latent relationships between artists. In the proposed method, a novel framework to construct a network is derived by collaboratively using social metadata and multi-modal features via canonical correlation analysis. Different from conventional methods that do not use multi-modal features, the proposed method can construct a network that can capture social metadata and multi-modal features, *i.e.*, a context-aware network. For effectively analyzing the context-aware network, a novel framework to realize popularity estimation of artists is developed based on network analysis. The proposed method enables effective utilization of the network structure by extracting node features via a node embedding algorithm. By constructing an estimator that can distinguish differences between the node features, the proposed method can archive accurate popularity estimation of artists. Experimental results using multiple real-world datasets that contain artists in various genres in Spotify, one of the largest MSS, are presented. Quantitative and qualitative evaluations show that our method is effective for both classifying and regressing the popularity.

INDEX TERMS Music, social network services, complex networks, prediction algorithms, classification algorithms.

I. INTRODUCTION

A large number of artists and users have utilized music streaming services (MSS) such as Spotify [1] and Apple Music [2] to upload artists' audio tracks [1] or listen to users' desired audio tracks [2]. In this situation, a system that can estimate popularity of artists can benefit artists to plan for improving their popularity. Specifically, artists can observe effects of uploading their audio tracks or information such as texts of a biography by simulating a change of popularity via the estimation system (see Fig. 1). Thus, there is a great demand for estimation of the popularity of artists in MSS [3]–[5].

To the best of our knowledge, however, there have been few works on automatic estimation of the popularity of artists in MSS, though some methods for simply formulating measurements of their popularity have been proposed [6], [7]. However, when we spread the focus to multimedia contents of social media, *e.g.*, images, videos, microblogs and audio tracks, many methods [8]–[25] have been proposed. For realizing popularity prediction, most of these methods analyze content features (*e.g.*, visual, textual and audio features) based on various approaches such as regression models [8]–[13] and deep-learning-based models [14]–[21]. If there are contents for which features are similar, these methods [8]–[21] predict that these contents have similar popularity. In other words, popularity predicted by these

The associate editor coordinating the review of this manuscript and approving it for publication was Michael Lyu.

[1] https://www.spotify.com/

[2] https://www.apple.com/music/

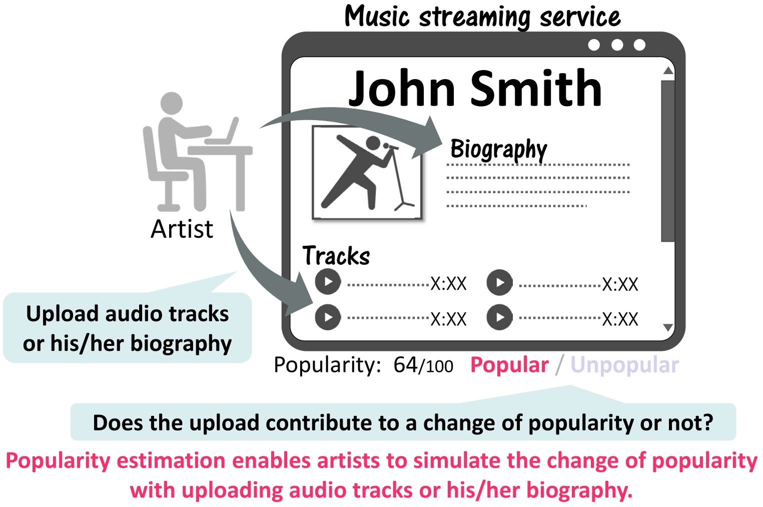

Music streaming service
John Smith
Biography
Tracks
X:XX
X:XX
X:XX
X:XX
Artist
Upload audio tracks or his/her biography
Popularity: 64/100 Popular / Unpopular
Does the upload contribute to a change of popularity or not?
Popularity estimation enables artists to simulate the change of popularity with uploading audio tracks or his/her biography.

FIGURE 1. An example that shows the usefulness of popularity estimation.

methods depends only on content features. However, if these methods are used to estimate the popularity of artists in MSS, there may be cases in which content features are only a part of the elements for estimating popularity. In MSS, there are abundant social metadata to associate artists with each other. It has been reported that social metadata in various social media are effective for popularity prediction since social metadata are attached by annotators after a semantic understanding [26]. Therefore, utilizing social metadata as well as content features is expected to be useful for realizing successful popularity estimation.

In recent years, methods that construct a network for which nodes are multimedia contents and links are defined by social metadata [22]–[25] have attracted much attention. Since a network structure based on social metadata can accurately represent relationships between multimedia contents in social media [27], these methods enable accurate popularity prediction. Here, artists in MSS have multi-modal features such as features of their audio tracks and their biographies, which characterize the artists. However, conventional methods [22]–[25] do not have extensibility for collaboratively using multi-modal features and social metadata. Thus, there is room for performance improvement by extending conventional methods to an alternative method that can collaboratively use them.

In this paper, we propose a novel method for a novel trial of estimating the popularity of artists based on network analysis. To the best of our knowledge, the method presented in this paper is the first attempt to automatically estimate the popularity of artists in MSS by collaborative use of multi-modal features and social metadata. The technical novelty of this paper is construction of a novel framework that can use uni-modal or multi-modal features commonly via analysis of a context-aware network, *i.e.*, a network that can capture not only social metadata but also multi-modal features. The technical contribution of this paper is to improve extensibility for using multi-modal features, which enables accurate analysis of latent relationships between artists. Specifically, we first construct the context-aware network by utilizing audio features of artists' audio tracks and textual features of artists' biographies and social metadata such as "related artists" via canonical correlation analysis (CCA) [28]. Thus, we can accurately represent relationships between artists. To effectively analyze the context-aware network, we construct an estimator using node features via node2vec [29]. Since the node features can represent characteristics of artists, the proposed method can accurately classify whether an artist is popular or not by using a classifier. Also, by using a regressor, we can estimate detail scores of the popularity (popularity scores). Finally, we verify the effectiveness of the proposed method by presenting results of experiments using multiple real-world datasets that contain artists in various genres such as pop, rock, classical, jazz and reggae in Spotify, which is one of the largest MSS.

II. RELATED WORKS

In this section, we explain recently published works on prediction of the popularity of multimedia contents in social media [8]–[25]. Analysis of content features in some methods is based on various regression models such as linear regression models [8]–[11] and decision tree regression models [11], [12]. Specifically, support vector regression (SVR) [30] is utilized for prediction of the popularity of videos by analyzing visual features [8] and for prediction of the popularity of images by analyzing visual and sentiment features [9]. Huang *et al.* [10] fused multiple regression models including k-nearest neighbors (kNN) and random forest [31] via SVR. Wang and Zhang [11] predicted the popularity of images by constructing ridge regression (RR) [32] and a gradient boosting regression tree (GBRT) [33]. Since these methods distinguish differences between content features,

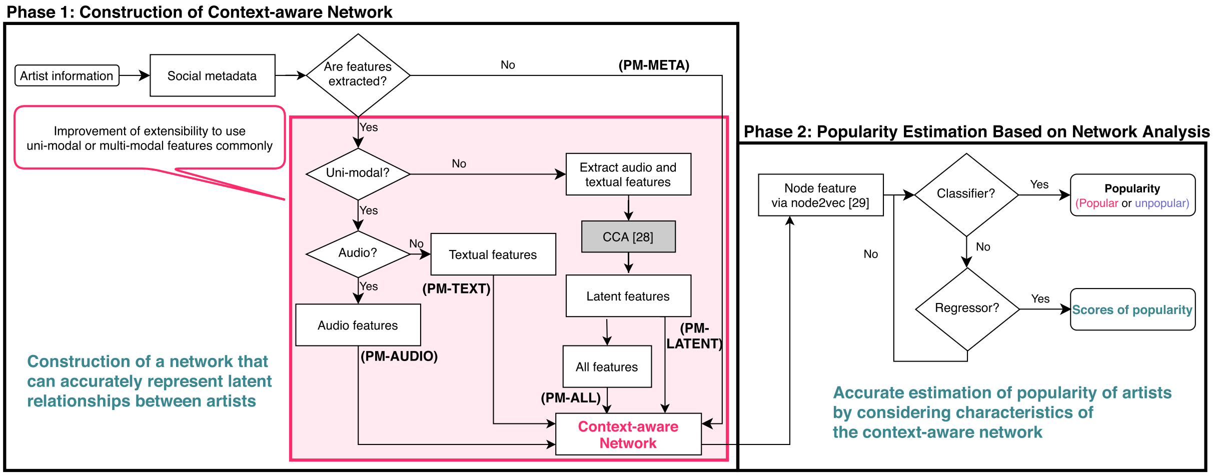

Phase 1: Construction of Context-aware Network
Artist information
Social metadata
Are features extracted?
No
(PM-META)
Yes
Improvement of extensibility to use uni-modal or multi-modal features commonly
Uni-modal?
No
Extract audio and textual features
Yes
CCA [28]
Audio?
No
Textual features
Yes
(PM-TEXT)
Latent features
Audio features
(PM-LATENT)
(PM-AUDIO)
All features
Construction of a network that can accurately represent latent relationships between artists
(PM-ALL)
Context-aware Network
Phase 2: Popularity Estimation Based on Network Analysis
Node feature via node2vec [29]
Classifier?
Yes
Popularity (Popular or unpopular)
No
No
Regressor?
Yes
Scores of popularity
Accurate estimation of popularity of artists by considering characteristics of the context-aware network

FIGURE 2. Overview of context-aware network analysis for popularity estimation of artists.

their performance relies on discriminant power of the content features.

To extract content features with high discriminant power, methods based on deep leaning such as convolutional neural networks (CNNs) [14]–[19] and recurrent neural networks (RNNs) [20], [21] have been utilized. Mao *et al.* [14] constructed a deep neural network that can predict the popularity of videos from features of attribute information of videos such as genres. Fontanini *et al.* [15] constructed a CNN based on visual and sentiment features of videos. In addition, some methods realize prediction of the popularity of images by constructing CNNs with visual features [16] and multiple features such as visual, textual and attribute features [17]. For prediction of the popularity of audio tracks, Yu *et al.* [18] used pre-trained music tagging CNN-based audio features and constructed a CNN that is suitable for popularity prediction. When constructing RNNs, long short-term memory (LSTM) [34] is often utilized for predicting the popularity of images [20] and videos [21]. In the above methods, predicted popularity depends only on content features. However, if we adopt these methods [8]–[21], successful popularity prediction cannot be realized in cases in which content features are only part of the elements for predicting popularity.

Since social metadata are effective for associating related contents with each other [26], [27], methods that construct a network for which nodes are multimedia contents and links are defined by social metadata [22]–[25] have been proposed. Specifically, the methods in the papers [22]–[24] realize popularity prediction by calculating page ranks [35] of networks based only on social metadata. On the other hand, Hong *et al.* [25] utilized not only social metadata but also textual features by constructing a network for prediction of the popularity of microblogs such as tweets.

Note that these methods [22]–[25] do not have a framework to capture multi-modal features. In our method, we improve extensibility for using multi-modal features and social metadata collaboratively by constructing a context-aware network.

III. CONTEXT-AWARE NETWORK ANALYSIS FOR POPULARITY ESTIMATION OF ARTISTS

An overview of the proposed method is shown in Fig. 2. As shown in Fig. 2, the proposed method consists of two phases: construction of a context-aware network (Section III-A) and popularity estimation based on network analysis (Section III-B).

A. CONSTRUCTION OF CONTEXT-AWARE NETWORK

In the proposed method, we construct a network $\mathcal{G}$ for which nodes are artists $m \in \mathcal{M}$ ($\mathcal{M}$ being a set of artists). Our framework to construct a context-aware network has extensibility for using only social metadata or using both social metadata and features of artists. Below, we explain the construction of the network in each case.

1) (PM-META): NETWORK CONSTRUCTION BASED ON SOCIAL METADATA

We define the existence of links on $\mathcal{G}$ based on social metadata. According to reports showing that social metadata are effective for associating similar contents with each other [27], [36], [37], we build links between m and n ($m, n \in \mathcal{M}$) if m is a "related artist" of n or vice versa. Since "related artists" can represent which artists are related to each other, we can effectively associate artists with each other. Therefore, we can construct a network that can capture social relationships between artists.

2) (PM-AUDIO) AND (PM-TEXT): NETWORK CONSTRUCTION BASED ON SOCIAL METADATA AND AUDIO OR TEXTUAL FEATURES

The proposed method can capture not only social metadata but also features of artists as follows. First, we define the existence of links in the same manner as (PM-META). Then, for each artist m, we extract an audio feature $\boldsymbol{v}_a^{(m)}$ from his/her audio tracks and a textual feature $\boldsymbol{v}_t^{(m)}$ from his/her biography. It is expected that characteristics of artists' audio tracks (*e.g.*, moods or genres) are related to their popularity. Also, characteristics of artists' biographies (*e.g.*, record of awards) can be related to their popularity. Therefore, such features are effective for estimating popularity. Finally, we define weights of links based on similarities of these features. Concretely, we calculate a link weight $w(m, n)$ between m and $n \in \mathcal{M}$, which links in $\mathcal{G}$, as follows:

$$w(m,n) = \left| \frac{\boldsymbol{v}^{(m)^{\mathrm{T}}} \boldsymbol{v}^{(n)}}{\|\boldsymbol{v}^{(m)}\| \ \|\boldsymbol{v}^{(n)}\|} \right|, \tag{1}$$

where $\boldsymbol{v}^{(m)} \in \{\boldsymbol{v}_a^{(m)}, \boldsymbol{v}_t^{(m)}\}$. It has been reported that calculation of similarities based on the above equation is effective for constructing a network for which nodes are contents on social media [38]–[40]. Therefore, we can capture characteristics of artists' audio tracks or biographies by constructing a network with either audio or textual features.

3) (PM-LATENT): NETWORK CONSTRUCTION BASED ON SOCIAL METADATA AND LATENT FEATURES

In the proposed method, we can collaboratively use social metadata and multi-modal features by calculating latent features. The latent features can capture latent relationships between multi-modal features by projecting audio and textual features into the same latent space via CCA. Specifically, we define the link existence and extract audio and textual features in the same manner as (PM-AUDIO) and (PM-TEXT). Here, a matrix $\boldsymbol{V}_\zeta = [\boldsymbol{v}_\zeta^{(1)} - \bar{\boldsymbol{v}}_\zeta, \boldsymbol{v}_\zeta^{(2)} - \bar{\boldsymbol{v}}_\zeta, \ldots, \boldsymbol{v}_\zeta^{(|\mathcal{M}|)} - \bar{\boldsymbol{v}}_\zeta]$, where $\bar{\boldsymbol{v}}_\zeta$ is the average vector of $\boldsymbol{v}_\zeta^{(m)}$, ($\zeta \in \{a, t\}$) is defined. Then we obtain projection matrices, which enable projections of $\boldsymbol{v}_\zeta^{(m)}$ into the same latent space via CCA. To calculate the latent features, audio and textual features must be directly compared with each other. The proposed method realizes direct comparison of them by adopting CCA. Specifically, we solve the following optimization problem.

$$\max_{\boldsymbol{u}_a, \boldsymbol{u}_t} \boldsymbol{u}_a^{\mathrm{T}} \boldsymbol{V}_{at} \boldsymbol{u}_t \ \text{ s.t. } \boldsymbol{u}_a^{\mathrm{T}} \boldsymbol{V}_{aa} \boldsymbol{u}_a = \boldsymbol{u}_t^{\mathrm{T}} \boldsymbol{V}_{tt} \boldsymbol{u}_t = \boldsymbol{1},$$

where $\boldsymbol{V}_{at}$ $\boldsymbol{V}_{aa}$ and $\boldsymbol{V}_{tt}$ are a variance-covariance matrix of audio and textual features, that of audio features and that of textual features, respectively. According to [28], this problem can be solved through the generalized eigenvalue problem as follows:

$$\begin{bmatrix} \boldsymbol{O} & \boldsymbol{V}_{at} \\ \boldsymbol{V}_{at}^{\mathrm{T}} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_a \\ \boldsymbol{u}_t \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{V}_{aa} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{V}_{tt} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_a \\ \boldsymbol{u}_t \end{bmatrix}. \tag{2}$$

By solving Eq. (2), we can obtain projection matrices $\boldsymbol{U}_\zeta$, which consists of $\boldsymbol{u}_\zeta$. By the following computation, we can transform $\boldsymbol{V}_\zeta$ to new feature matrices $\hat{\boldsymbol{V}}_\zeta = [\hat{\boldsymbol{v}}_\zeta^{(1)}, \hat{\boldsymbol{v}}_\zeta^{(2)}, \ldots, \hat{\boldsymbol{v}}_\zeta^{|\mathcal{M}|}]$, which can be directly compared with each other:

$$\hat{\boldsymbol{V}}_\zeta = \boldsymbol{U}_\zeta^{\mathrm{T}} \boldsymbol{V}_\zeta.$$

As a result, we can calculate latent features $\boldsymbol{v}_l^{(m)}$ by the following equation:

$$\boldsymbol{v}_l^{(m)} = [(\hat{\boldsymbol{v}}_a^{(m)})^{\mathrm{T}}, (\hat{\boldsymbol{v}}_t^{(m)})^{\mathrm{T}}]^{\mathrm{T}}. \tag{3}$$

By using the latent features, we can consider latent characteristics of both audio and textual features. Finally, we define link weights by Eq. (1). Therefore, the constructed network can represent latent relationships between artists.

4) (PM-ALL): NETWORK CONSTRUCTION BASED ON SOCIAL METADATA AND ALL OF THE AUDIO, TEXTUAL AND LATENT FEATURES

To consider all of the audio, textual and latent features, we construct a network based on social metadata and audio, textual and latent features. First, we define the link existence in the same manner as (PM-META) and extract audio, textual features by the same computation as that for (PM-AUDIO), (PM-TEXT) and (PM-LATENT), respectively. Then we calculate features by the following computation:

$$\boldsymbol{v}_{\mathrm{all}}^{(m)} = [(\boldsymbol{v}_a^{(m)})^{\mathrm{T}}, (\boldsymbol{v}_t^{(m)})^{\mathrm{T}}, (\boldsymbol{v}_l^{(m)})^{\mathrm{T}}]^{\mathrm{T}}. \tag{4}$$

By Eq. (4), we can consider all of the audio, textual and visual features. It is expected that these combined features have stronger discriminant power than that of only audio, textual or latent features. Therefore, by using the features, accurate popularity estimation can be realized. Consequently, we can construct a network that can capture both multi-modal features and social metadata by weighting links via Eq. (1).

B. POPULARITY ESTIMATION BASED ON NETWORK ANALYSIS

We first calculate node features $\tilde{\boldsymbol{u}}^{(m)}$ by applying node2vec [29] to the obtained network $\mathcal{G}$. In the node2vec algorithm, we search for a network neighborhood of nodes by iteratively performing a random walk [41]. Then we optimize the objective function, which maximizes the log-probability of observing the network neighborhood via stochastic gradient descent (SGD). Thus, we can know not only "which node is a neighbor node of each node" but also "which node has a similar structure on $\mathcal{G}$." By using $\tilde{\boldsymbol{u}}^{(m)}$, we construct an estimator that can consider characteristics of context-aware network. If we construct a classifier for which positive samples are popular artists and negative samples are unpopular artists, we can estimate whether an artist is popular or not by inputting $\tilde{\boldsymbol{u}}^{(m_t)}$ (m_t being an artist to be estimated). This is beneficial for artists who desire to know their approximate popularity. On the other hand, if we construct a regressor for which training samples are known popularity scores, we can estimate popularity scores. This is also beneficial for artists who desire to know their popularity in more detail.

To adopt network analysis, a network including sufficient information on artists must be constructed. However, such a network, which can capture multi-modal features and social metadata, has not been constructed so far. In the proposed method, we can construct a context-aware network via CCA.

TABLE 1. Details of datasets used in the experiments.

Dataset Name	Name of Seed Artists	Genre Name of Seed Artists	Num. of Nodes	Num. of Links
Dataset #1	Lady Gaga	Pop	2,992	14,985
Dataset #2	The Beatles	Rock	3,000	17,660
Dataset #3	Andrea Bocelli	Classical	3,000	15,771
Dataset #4	Louis Armstrong	Jazz	2,999	20,619
Dataset #5	Jimmy Cliff	Reggae	3,000	17,232

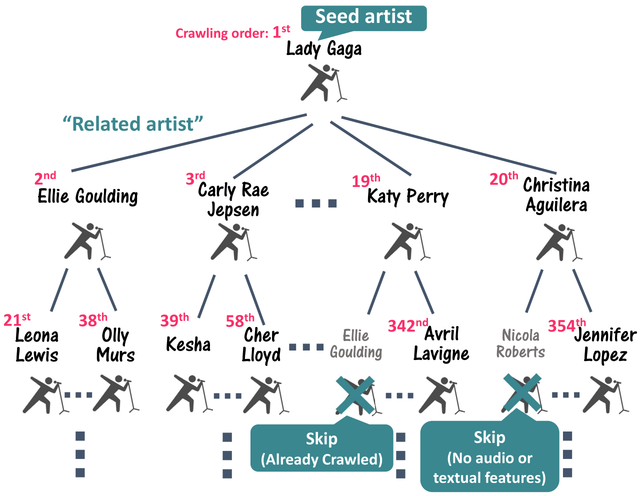


FIGURE 3. Details of the crawling procedure to search for artists in dataset #1.

As a result, the proposed method can accurately estimate artists' popularity.

IV. EXPERIMENTAL RESULTS

In this section, we show the effectiveness of the proposed method by estimating the popularity of artists in Spotify through experiment using both a classifier and a regressor.

A. SETTINGS

In the experiment, we constructed datasets by collecting information about real-world artists and their audio tracks via Spotify API.[3] Specifically, we first defined artists in Table 1 as "seed artists," *i.e.*, the first artists for crawling. For each seed artist, we crawled "popularity[4]" on Spotify and "related artists," *i.e.*, metadata for associating artists related to each other. Then, by repeating the crawling of popularity and related artists for each artist of related artists, we collected information on a maximum of 3,000 artists. The size of dataset was set to approximately the same or more than that used in conventional methods [6], [7]. When crawling the information, we only crawled artists who have both audio and textual features. Details of the crawling procedure to search for artists are shown in Fig. 3. In Dataset #1, we first crawled related artists and other information of "Lady Gaga". Next, we crawled the same information of related artists of "Lady Gaga", *e.g.*, "Ellie Goulding", "Carly Rae Jepsen", "Katy Perry" and "Christina Aguilera". Further crawling is performed by repeating this procedure. This procedure corresponds to the breadth first search.

TABLE 2. Summary of audio features provided by Spotify API. Description is from a paper [42].

Feature Name	Description
Acousticness	A confidence measure of whether the track is acoustic.
Danceability	Describes how suitable a track is for dancing based on a combination of musical elements.
Duration in ms	Duration of the track in milliseconds.
Energy	Represents a perceptual measure of intensity and activity. Energetic tracks feel fast, loud, and noisy.
Instrumentalness	Predicts whether a track contains no vocals. Rap or spoken word tracks are clearly 'vocal'.
Key	The key the track is in. Integers map to pitches using standard Pitch Class notation.
Liveness	Detects the presence of an audience in the recording.
Loudness	The overall average loudness of a track in dB.
Mode	Indicates the modality (major or minor) of a track.
Speechiness	Detects the presence of spoken words in a track. The more exclusively speech-like the recording is (*e.g.*, talk show, audio book, poetry), the closer to 1.0 is the attribute value becomes.
Tempo	The overall estimated tempo of a track in beats per minute (BPM).
Valence	Describes the musical positiveness. Tracks with high valence sound happier, while tracks with low valence sound more negative and sad.

Audio features were extracted from "top tracks" of each artist, which were provided in Spotify API. Specifically, for each artist, we collected top tracks as much as possible (max of 10 tracks). Then we calculated 24-dimensional features for which elements are means and standard deviations of the features shown in Table 2. Since it has been reported that

[3] https://developer.spotify.com/

[4] "Popularity" denotes an integer value in [0,100]. The higher these values of artists are, the more popular they are.

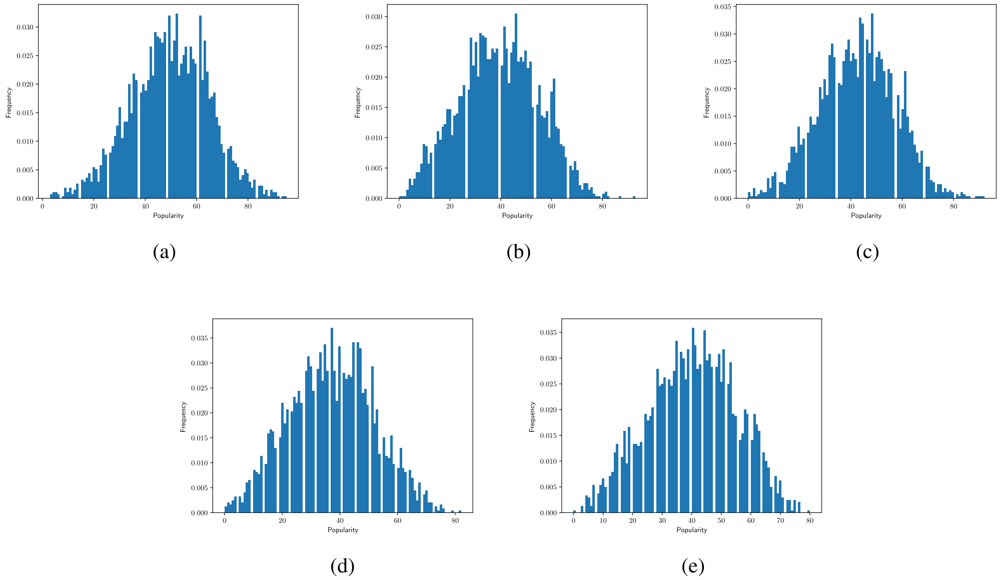


FIGURE 4. Histogram of the popularity of all artists in each dataset. (a): Dataset #1. (b): Dataset #2. (c): Dataset #3. (d): Dataset #4. (e): Dataset #5.

audio features in Table 2 can represent basic information of audio tracks [42], these features can be effective for estimation of the popularity of artists. In addition, we applied doc2vec [43] to the biography of each artist on the Web page of Spotify and obtained 100-dimensional textual features. Since doc2vec can consider the order in which words appear and learn semantic representation, it is expected that these textual features can represent each artist's biography.

Figure 4 shows a histogram of the popularity of all artists in each dataset. In Fig. 4, we can see that the histogram has an almost normal distribution. Thus, in the experiment, we defined the mean value of all artists' popularity as a criterion that decides whether an artist is popular or not. In other words, we defined artists who have the same popularity as the mean value or higher popularity than the mean value as popular artists. We also defined artists who have lower popularity than the mean value as unpopular artists. In this way, we provided the ground truth (GT) of the popularity for each artist for classification. Also, we set the crawled popularity as the GT for regression. It should be noted that popularity of artists[5] is calculated mathematically on the basis of popularity of their audio tracks[6] and the value is not updated in real time. Therefore, we can consider that calculation of the popularity is consistent.

To evaluate whether popularity provided by Spotify corresponds to human perception or not, we conducted a subject experiment. First, we randomly extracted some artists from each dataset and ordered them in descending order of their popularity. In the experiment, we set the number of extracted artists to 4 according to the report showing that people can only grasp about four chunks in short-term memory tasks [44]. We repeated to extract rankings of artists for 4 times for each dataset, *i.e.*, for 20 times in total. Then 14 subjects (11 males and 3 females, ages 22-25) evaluated whether they considered each ranking of artists is appropriate or not and gave a score from 1 to 4 (1: Not appropriate, 2: Not appropriate much, 3: A little appropriate, 4: Appropriate). Table 3 shows the results of the subject experiment. Note that we show the mean scores for the rankings that are extracted from each dataset. "Mean" shows the mean of the calculated mean scores of all datasets. From the result, we can see that the mean reaches 3.40 in "Mean". Therefore, it is confirmed that the subjects mostly considered that the popularity provided in Spotify is appropriate. Above all, we can consider that popularity provided by Spotify corresponds to human perception and is suitable for GT of popularity estimation.

TABLE 3. Mean scores for each dataset.

Dataset No.	#1	#2	#3	#4	#5	Mean
	3.50	3.48	3.41	3.39	3.20	3.40

Furthermore, we adopted five-fold cross validation as the verification method for fair evaluations. Specifically, we split artists into training and test artists so that the percentages of training and test artists become 80% and 20%, respectively. Then we constructed a classifier and a regressor by using training artists. According to a report showing that a support vector machine (SVM) [45] is effective for node classification based on node features [46], we adopted an SVM as a classifier. In the same manner, we adopted support vector regression (SVR) as a regressor. Note that the kernel function in the SVM and SVR was the Gaussian kernel with parameters determined through a grid search [47]. Moreover, when

[5] https://developer.spotify.com/documentation/web-api/reference/artists/get-artist/ (accessed July 2, 2019).

[6] https://developer.spotify.com/documentation/web-api/reference/tracks/get-track/ (accessed July 2, 2019).

calculating latent features, we obtained projection matrices by Eq. (2).

B. QUANTITATIVE EVALUATIONS

To evaluate the effectiveness of the proposed method, we compared the proposed method (PM) with the following reference methods (RMs1-5).

(RM1) This is a method based on a recently published paper [11]. According to the paper [11], (RM1) constructs RR and GBRT by using features that are calculated in the same manner as that in Eq. (4). For classification, (RM1) constructs a ridge classifier and a gradient boosting decision tree (GBDT). We denote (RM1) with the RR/ridge classifier and the GBRT/GBDT by (RM1-R) and (RM1-G), respectively.

(RM2) This is a method that utilizes not social metadata but multi-modal features. Note that this method does not construct a network. Specifically, this method constructs an SVM by using features that are calculated in the same manner as that in Eq. (4) and sets parameters of the SVM in the same manner as that of (PM).

(RM3) This is a method that utilizes only audio features. Specifically, this method constructs a classifier by using audio features in the same manner as that of (RM2).

(RM4) This is a method that utilizes only textual features. Specifically, this method performs estimation in the same manner as that of (RM2).

(RM5) This is a method that utilizes only latent features. Specifically, this method calculates latent features in Eq. (3) and performs estimation in the same manner as that of (RM2).

Note that (RMs2-5) follow the idea that uses only content features for constructing regression models using support vectors such as those in the conventional methods in the papers [8], [9].

For classification, we used accuracy, precision, recall and F-measure of those estimated as popular artists. We also used precision$_u$, recall$_u$ and F-measure$_u$ of those estimated as unpopular artists and Matthew's correlation coefficients (MCCs). These evaluation metrics are defined as follows:

$$\mathrm{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}},$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$

$$\mathrm{F-measure} = \frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}},$$

$$\mathrm{Precision}_u = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}},$$

$$\mathrm{Recall}_u = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},$$

$$\mathrm{F-measure}_u = \frac{2\times\mathrm{Precision}_u\times\mathrm{Recall}_u}{\mathrm{Precision}_u+\mathrm{Recall}_u},$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}},$$

where TP, TN, FP and FN are defined as follows.

TP	: Number of correctly estimated popular artists,
FN	: Number of incorrectly estimated popular artists,
FP	: Number of incorrectly estimated unpopular artists,
TN	: Number of correctly estimated unpopular artists.

For regression, we used mean absolute error (MAE) and mean absolute percentage error (MAPE), which are defined as follows:

$$\mathrm{MAE} = \frac{1}{N_{\mathrm{test}}}\sum_{i=1}^{N_{\mathrm{test}}}|s_{\mathrm{est}}^{(i)} - s_{\mathrm{GT}}^{(i)}|,$$

$$\mathrm{MAPE} = \frac{1}{N_{\mathrm{test}}}\sum_{i=1}^{N_{\mathrm{test}}}\left|\frac{s_{\mathrm{est}}^{(i)} - s_{\mathrm{GT}}^{(i)}+1}{s_{\mathrm{GT}}^{(i)}+1}\right|,$$

where N_{test}, $s_{\mathrm{est}}^{(i)}$ and $s_{\mathrm{GT}}^{(i)}$ are the number of test artists, the i-th artist's estimated popularity score, and the i-th artist's GT, respectively. Note that we added 1 to both of the denominator and numerator of MAPE to avoid the situation in which the denominator is 0.

Tables 4-8 show classification results for each dataset, and Table 9 shows mean values of classification results for all datasets. Also, Tables 10 and 11 show regression results for all datasets. From Tables 9-11, we can see that all of the PMs have mostly outperformed (RMs1-5) for all evaluation metrics of both classification and regression. From the results obtained by all of the PMs, (RM1-R) and (RM1-G), we can see that all of the PMs have greatly outperformed the other methods for all evaluation measures and for all datasets. Also, we can confirm that all of the PMs outperformed (RM2), which utilizes not network representation but multi-modal features. From the results, the effectiveness of constructing a network representation based on social metadata can be verified. A comparison of all of the PMs with (RMs3-5) shows the effectiveness of utilizing multi-modal features and constructing the network representation. As shown in the mean value of all evaluation metrics in Tables 9-11, (PM-ALL) mostly outperformed (PM-META), (PM-AUDIO), (PM-TEXT) and (PM-LATENT). Therefore, it is thought that there are many cases in which estimation of the latent relationships between multi-modal features via CCA is effective for popularity estimation. As a result, we can confirm that all of the PMs are effective for classification and regression of the popularity of artists.

C. QUALITATIVE EVALUATION

In 1), we discuss the effectiveness of constructing the proposed network by visualization. In 2) and 3), we discuss the effectiveness and limitations of the proposed method in detail by showing examples of classification and regression results.

TABLE 4. Classification result for dataset #1.

	(PM-ALL)	(PM-AUDIO)	(PM-TEXT)	(PM-LATENT)	(PM-META)	(RM1-R) [11]	(RM1-G) [11]	(RM2)	(RM3)	(RM4)	(RM5)
Accuracy	**0.792**	0.787	0.783	0.776	**0.792**	0.682	0.670	0.686	0.647	0.644	0.660
Precision	0.791	**0.796**	0.791	0.780	0.793	0.682	0.671	0.691	0.623	0.685	0.653
Recall	**0.801**	0.789	0.780	0.779	0.794	0.700	0.684	0.689	0.769	0.551	0.710
F-measure	**0.796**	0.788	0.784	0.779	0.793	0.690	0.677	0.689	0.688	0.610	0.679
Precision$_u$	**0.794**	0.780	0.778	0.774	0.780	0.685	0.670	0.683	0.688	0.617	0.674
Recall$_u$	0.784	**0.794**	0.788	0.775	0.789	0.666	0.656	0.684	0.523	0.740	0.611
F-measure$_u$	**0.789**	0.787	0.783	0.774	0.784	0.675	0.663	0.683	0.594	0.673	0.641
MCC	**0.585**	0.575	0.569	0.554	**0.585**	0.367	0.340	0.374	0.302	0.297	0.324

TABLE 5. Classification result for dataset #2.

	(PM-ALL)	(PM-AUDIO)	(PM-TEXT)	(PM-LATENT)	(PM-META)	(RM1-R) [11]	(RM1-G) [11]	(RM2)	(RM3)	(RM4)	(RM5)
Accuracy	**0.834**	0.824	0.824	0.817	0.825	0.679	0.676	0.686	0.622	0.657	0.671
Precision	**0.835**	0.827	0.828	0.818	0.833	0.700	0.677	0.705	0.604	0.695	0.682
Recall	**0.829**	0.817	0.816	0.812	0.810	0.620	0.667	0.633	0.694	0.558	0.631
F-measure	**0.832**	0.821	0.821	0.815	0.821	0.657	0.672	0.666	0.645	0.617	0.655
Precision$_u$	**0.833**	0.822	0.821	0.816	0.818	0.664	0.670	0.672	0.646	0.635	0.662
Recall$_u$	**0.839**	0.831	0.832	0.822	**0.839**	0.738	0.686	0.738	0.552	0.755	0.711
F-measure$_u$	**0.836**	0.826	0.826	0.819	0.828	0.699	0.678	0.703	0.595	0.690	0.686
MCC	**0.668**	0.648	0.648	0.634	0.651	0.361	0.353	0.374	0.248	0.322	0.343

TABLE 6. Classification result for dataset #3.

	(PM-ALL)	(PM-AUDIO)	(PM-TEXT)	(PM-LATENT)	(PM-META)	(RM1-R) [11]	(RM1-G) [11]	(RM2)	(RM3)	(RM4)	(RM5)
Accuracy	**0.813**	0.805	0.801	0.802	0.810	0.667	0.676	0.680	0.642	0.649	0.677
Precision	**0.815**	0.807	0.810	0.810	0.811	0.667	0.680	0.698	0.642	0.683	0.675
Recall	**0.823**	0.817	0.802	0.802	0.821	0.701	0.698	0.698	0.685	0.593	0.714
F-measure	**0.818**	0.811	0.805	0.806	0.816	0.682	0.688	0.691	0.663	0.634	0.693
Precision$_u$	**0.811**	0.804	0.793	0.793	0.808	0.668	0.673	0.676	0.642	0.624	0.680
Recall$_u$	**0.803**	0.794	0.802	0.802	0.799	0.633	0.655	0.663	0.597	0.710	0.639
F-measure$_u$	**0.807**	0.799	0.797	0.797	0.803	0.650	0.664	0.669	0.619	0.664	0.659
MCC	**0.626**	0.611	0.604	0.604	0.620	0.334	0.353	0.361	0.283	0.305	0.354

TABLE 7. Classification result for dataset #4.

	(PM-ALL)	(PM-AUDIO)	(PM-TEXT)	(PM-LATENT)	(PM-META)	(RM1-R) [11]	(RM1-G) [11]	(RM2)	(RM3)	(RM4)	(RM5)
Accuracy	0.808	**0.810**	0.804	0.799	0.803	0.648	0.660	0.679	0.622	0.646	0.663
Precision	0.810	**0.813**	0.802	0.804	0.805	0.647	0.655	0.678	0.610	0.673	0.663
Recall	0.792	0.791	**0.793**	0.776	0.784	0.602	0.635	0.647	0.617	0.532	0.631
F-measure	**0.801**	**0.801**	0.797	0.790	0.794	0.623	0.644	0.661	0.612	0.592	0.645
Precision$_u$	**0.807**	**0.807**	0.806	0.794	0.800	0.648	0.664	0.680	0.635	0.631	0.665
Recall$_u$	0.824	**0.828**	0.815	0.821	0.821	0.690	0.683	0.711	0.628	0.756	0.694
F-measure$_u$	0.815	**0.817**	0.810	0.807	0.810	0.668	0.673	0.695	0.631	0.688	0.679
MCC	0.616	**0.620**	0.608	0.598	0.606	0.294	0.318	0.358	0.245	0.296	0.327

TABLE 8. Classification result for dataset #5.

	(PM-ALL)	(PM-AUDIO)	(PM-TEXT)	(PM-LATENT)	(PM-META)	(RM1-R) [11]	(RM1-G) [11]	(RM2)	(RM3)	(RM4)	(RM5)
Accuracy	0.820	**0.824**	0.818	0.818	0.813	0.671	0.685	0.700	0.632	0.655	0.687
Precision	0.817	**0.818**	0.817	**0.818**	0.816	0.683	0.689	0.709	0.629	0.702	0.688
Recall	0.823	**0.831**	0.818	0.816	0.807	0.634	0.669	0.674	0.641	0.536	0.680
F-measure	0.820	**0.825**	0.817	0.817	0.811	0.658	0.679	0.691	0.633	0.607	0.684
Precision$_u$	0.823	**0.829**	0.819	0.817	0.810	0.661	0.681	0.692	0.639	0.627	0.686
Recall$_u$	0.816	0.816	0.817	**0.819**	**0.819**	0.708	0.700	0.726	0.626	0.774	0.695
F-measure$_u$	0.819	**0.822**	0.818	0.818	0.814	0.684	0.690	0.709	0.632	0.693	0.690
MCC	0.639	**0.648**	0.635	0.636	0.625	0.343	0.370	0.400	0.267	0.319	0.375

1) VISUALIZATION OF CONTEXT-AWARE NETWORK

To qualitatively evaluate the effectiveness of constructing the network, we visualized the network in Fig. 5. For easy-to-see visualization, we utilized the spring model ForceAtlas2 [48], which is provided in the visualization tool gephi [49]. Note that ForceAtlas2 locates similar nodes in the neighborhood and dissimilar nodes at a distance. In other words, ForceAtlas2 can visualize a group of similar nodes in a network as a cluster. In Fig. 5(a), we show an overview of the network with Dataset #1. From this figure, we can see that

TABLE 9. Classification results for all datasets. Elements are mean values of all datasets.

	(PM-ALL)	(PM-AUDIO)	(PM-TEXT)	(PM-LATENT)	(PM-META)	(RM1-R) [11]	(RM1-G) [11]	(RM2)	(RM3)	(RM4)	(RM5)
Accuracy	**0.813**	0.810	0.806	0.803	0.804	0.669	0.673	0.686	0.633	0.650	0.672
Precision	**0.814**	0.812	0.809	0.806	0.811	0.676	0.674	0.694	0.622	0.688	0.672
Recall	**0.814**	0.809	0.802	0.797	0.803	0.651	0.671	0.668	0.681	0.554	0.673
F-measure	**0.813**	0.809	0.805	0.801	0.807	0.662	0.672	0.680	0.648	0.612	0.671
Precision$_u$	**0.814**	0.808	0.803	0.799	0.803	0.665	0.672	0.681	0.650	0.627	0.673
Recall$_u$	**0.813**	**0.813**	0.811	0.808	**0.813**	0.687	0.676	0.704	0.585	0.747	0.670
F-measure$_u$	**0.813**	0.810	0.807	0.803	0.808	0.675	0.674	0.692	0.614	0.682	0.671
MCC	**0.627**	0.620	0.613	0.605	0.617	0.340	0.347	0.373	0.269	0.308	0.345

TABLE 10. MAE of regression results for all datasets.

	(PM-ALL)	(PM-AUDIO)	(PM-TEXT)	(PM-LATENT)	(PM-META)	(RM1-R) [11]	(RM1-G) [11]	(RM2)	(RM3)	(RM4)	(RM5)
Dataset #1	**6.71**	6.78	6.78	6.98	6.77	10.2	10.1	10.0	10.4	10.8	11.0
Dataset #2	6.43	6.42	6.41	6.55	**6.33**	10.5	10.3	10.2	10.5	11.7	10.7
Dataset #3	6.65	**6.60**	6.62	6.73	6.64	10.4	10.0	10.1	10.4	10.7	10.8
Dataset #4	**6.71**	6.74	6.72	6.89	6.76	10.5	10.2	10.1	10.6	11.1	10.6
Dataset #5	6.57	**6.49**	6.56	6.61	6.57	10.3	9.88	9.88	10.2	10.1	10.6
Mean	**6.61**	**6.61**	6.62	6.75	6.62	10.4	10.1	10.1	10.4	11.1	10.7

TABLE 11. MAPE of regression results for all datasets.

	(PM-ALL)	(PM-AUDIO)	(PM-TEXT)	(PM-LATENT)	(PM-META)	(RM1-R) [11]	(RM1-G) [11]	(RM2)	(RM3)	(RM4)	(RM5)
Dataset #1	**17.2**	17.5	17.5	17.9	17.5	26.8	26.6	26.5	27.6	28.7	29.0
Dataset #2	24.9	25.3	25.0	25.3	**24.7**	39.4	39.1	38.3	39.9	45.4	40.0
Dataset #3	20.7	20.8	**20.4**	20.8	20.7	36.7	35.7	36.0	37.2	38.6	38.3
Dataset #4	27.2	27.3	27.5	**27.1**	27.2	44.7	43.6	43.3	45.8	48.8	46.1
Dataset #5	22.2	**21.9**	22.3	22.2	22.0	37.2	36.1	35.8	37.2	40.4	38.5
Mean	**22.4**	22.6	22.5	22.7	22.5	37.0	36.2	36.0	37.5	40.4	38.4

most of the unpopular artists' nodes are distant from the center of the network. Next, we excerpted a part of the network in Fig. 5(b) to confirm the details of the network. To improve visibility of clusters, we laid out the network in Fig. 5(b) again and showed the result in Fig. 5(c). In Fig. 5(c), the largest cluster of the network consists of artists who mainly flourished in the 1970s to 1990s. The upper left and upper right clusters in Fig. 5(c) correspond to groups of country artists in the U.S. and trumpeters, respectively. Therefore, we can confirm that latent characteristics of artists that are complexly integrated with multiple elements such as genres, activity areas and active periods can be represented by constructing the network.

2) DISCUSSION OF EFFECTIVENESS VIA SUCCESSFUL EXAMPLES

Figures 6(a)-(c) show the successful examples of estimation results. Below, we discuss effectiveness of the proposed method via each example.

a: COMPARISON OF (PM-ALL) WITH (RM2)

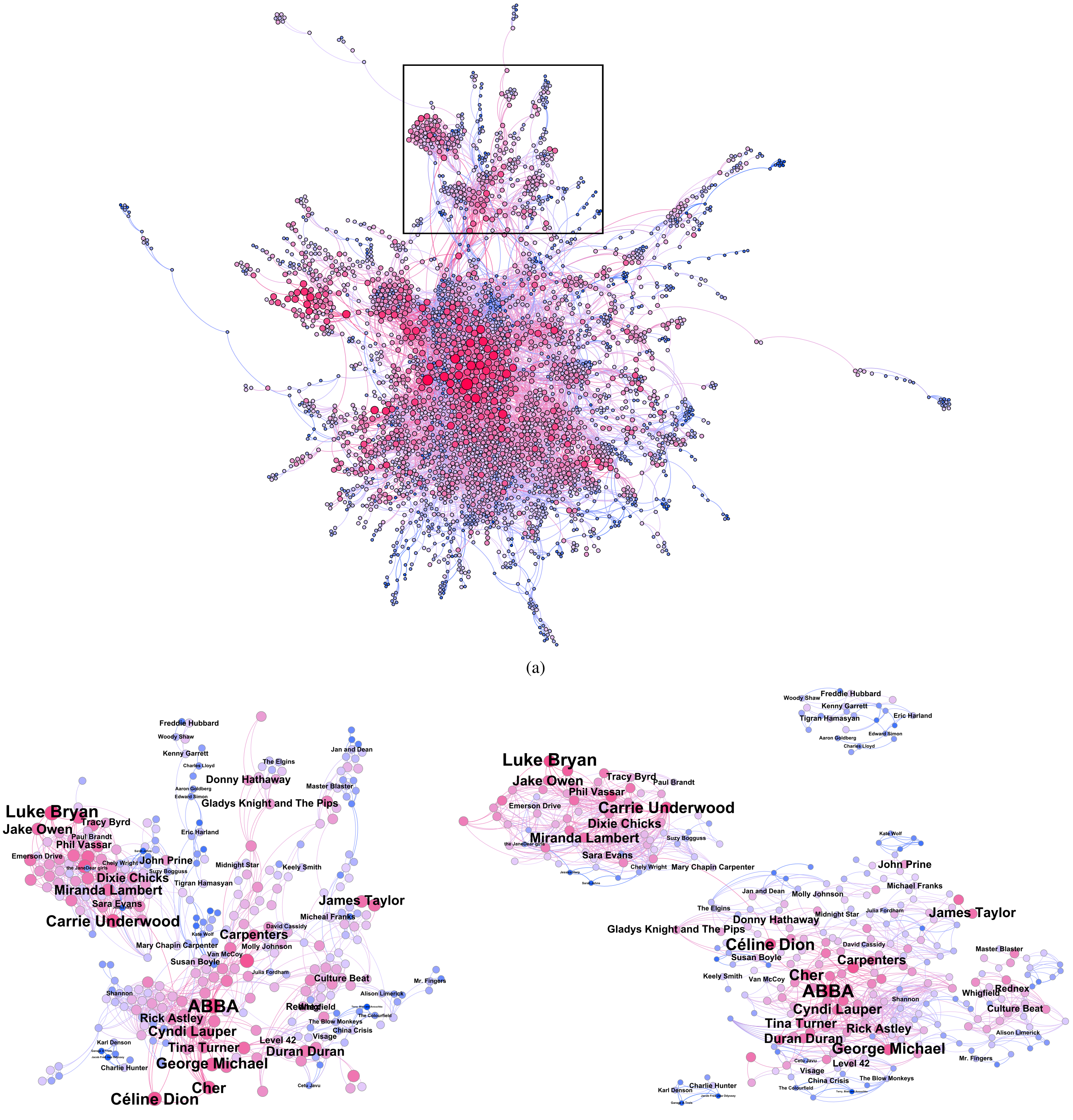
Figure 6(a) shows the result in the case where (PM-ALL) can successfully classify and regress the popularity and (RM2) cannot do. The artist "Cyndi Lauper" is placed in the center of the largest cluster in Fig. 5(c). Since this cluster consists of artists who mainly flourished in the 1970s to 1990s, relationships between these artists are represented in the proposed network. Thus, as described in 1) of this section, it can be confirmed that the context-aware network can represent latent characteristics of artists.

b: COMPARISON OF (PM-ALL) WITH (PM-AUDIO) AND (PM-TEXT)

Figure 6(b) shows the result when we compared (PM-ALL) with (PM-AUDIO). The biography of the artist "Sub Focus" includes the history of received awards, which may be information related to his popularity. Thus, it is confirmed that (PM-ALL) is more effective than (PM-AUDIO) when there is sufficient information in the artists' biography. Moreover, Fig. 6(c) shows the result when we compared (PM-ALL) with (PM-TEXT). In this figure, the biography of the artist "Poppy" is short and the method cannot obtain detail information of the artist. Therefore, it is thought that accurate estimation is difficult for (PM-TEXT), which only uses textual features. On the other hand, most audio tracks of the artist have features of electronic pops, one of the recently popular genres. Thus, it is confirmed that (PM-ALL) can accurately estimate the popularity by considering the audio features of the audio tracks.

3) DISCUSSION OF LIMITATION VIA UNSUCCESSFUL EXAMPLES

Figures 6(d)-(f) show the unsuccessful examples of estimation results. Below, we discuss limitations of the proposed method via each example.



(a)

(b)

(c)

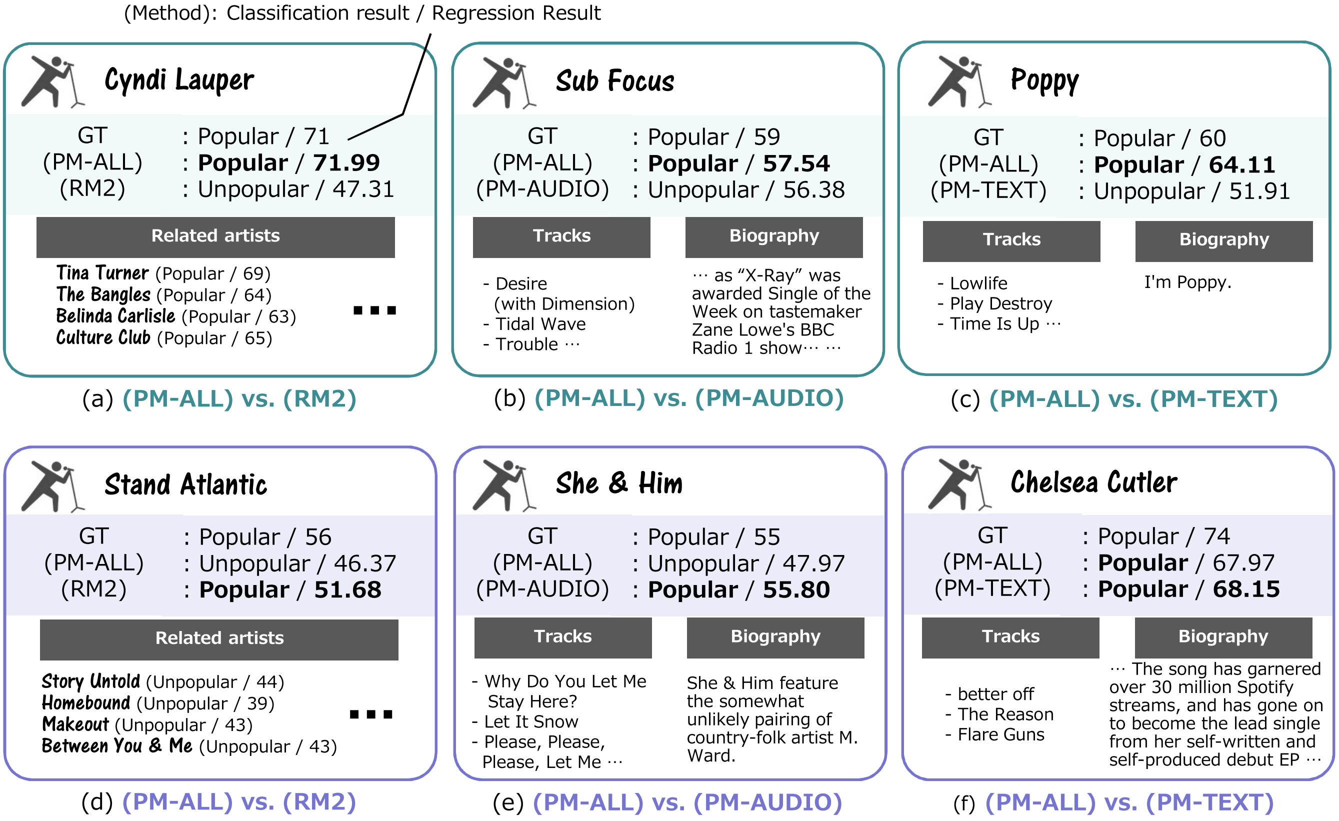
FIGURE 5. **Visualization of the constructed network with Dataset #1. Red and blue nodes correspond to popular and unpopular artists, respectively. As the node color changes from blue to red, popularity rises. (a): Overview of the constructed network with Dataset #1. (b): Excerpt network within the black frame in Fig. 5(a). (c): Network after re-layout of the network in Fig. 5(b) for improving visibility of clusters.**

a: COMPARISON OF (PM-ALL) WITH (RM2)

In Fig. 6(d), it is confirmed that social metadata worked better than multi-modal features. (PM-ALL) may mistake the estimation results by strongly considering the relationships between ''related artists'', whose popularity is different from that of the target artist. Thus, it is expected that the performance of the proposed method will be improved if we introduce a framework to control effects of social metadata and multi-modal features.

b: COMPARISON OF (PM-ALL) WITH (PM-AUDIO) AND (PM-TEXT)

Next, we compared (PM-ALL) with (PM-AUDIO) and (PM-TEXT) to clarify the limitation of using multi-modal


(Method): Classification result / Regression Result
Cyndi Lauper
GT : Popular / 71
(PM-ALL) : Popular / 71.99
(RM2) : Unpopular / 47.31
Related artists
Tina Turner (Popular / 69)
The Bangles (Popular / 64)
Belinda Carlisle (Popular / 63)
Culture Club (Popular / 65)
...
Sub Focus
GT : Popular / 59
(PM-ALL) : Popular / 57.54
(PM-AUDIO) : Unpopular / 56.38
Tracks
- Desire (with Dimension)
- Tidal Wave
- Trouble ⋯
Biography
⋯ as "X-Ray" was awarded Single of the Week on tastemaker Zane Lowe's BBC Radio 1 show⋯ ⋯
Poppy
GT : Popular / 60
(PM-ALL) : Popular / 64.11
(PM-TEXT) : Unpopular / 51.91
Tracks
- Lowlife
- Play Destroy
- Time Is Up ⋯
Biography
I'm Poppy.
Stand Atlantic
GT : Popular / 56
(PM-ALL) : Unpopular / 46.37
(RM2) : Popular / 51.68
Related artists
Story Untold (Unpopular / 44)
Homebound (Unpopular / 39)
Makeout (Unpopular / 43)
Between You & Me (Unpopular / 43)
...
She & Him
GT : Popular / 55
(PM-ALL) : Unpopular / 47.97
(PM-AUDIO) : Popular / 55.80
Tracks
- Why Do You Let Me Stay Here?
- Let It Snow
- Please, Please, Please, Let Me ⋯
Biography
She & Him feature the somewhat unlikely pairing of country-folk artist M. Ward.
Chelsea Cutler
GT : Popular / 74
(PM-ALL) : Popular / 67.97
(PM-TEXT) : Popular / 68.15
Tracks
- better off
- The Reason
- Flare Guns
Biography
⋯ The song has garnered over 30 million Spotify streams, and has gone on to become the lead single from her self-written and self-produced debut EP ⋯

(a) (PM-ALL) vs. (RM2)
(b) (PM-ALL) vs. (PM-AUDIO)
(c) (PM-ALL) vs. (PM-TEXT)
(d) (PM-ALL) vs. (RM2)
(e) (PM-ALL) vs. (PM-AUDIO)
(f) (PM-ALL) vs. (PM-TEXT)

FIGURE 6. Examples of classification/regression results in dataset #1. Figs. 6(a)-(c) show the successful results and Figs. 6(d)-(f) show the unsuccessful results.

features. In Fig. 6(e), we can confirm that there is less information related to the popularity in the biography. Thus, it is thought that the performance of (PM-ALL), which uses both audio and textual features, can be affected by the existence of information effective for popularity estimation. In Fig. 6(f), we can see that the artist "Chelsea Cutler" has become recently popular in Spotify according to the biography. Therefore, the biography may be highly correlated to the popularity and strongly effective for popularity estimation. It is thought that this is because (PM-TEXT) outperformed (PM-ALL) in Fig. 6(f). Note that the quality and quantity of information that is provided in MSS differs with each artist. To outcome the difficulty, we will introduce a framework to improve the discriminant power of both features or select optimal features for estimation of the popularity of each artist as a future work.

V. FUTURE WORK

Future work of this paper is described in this section. As described in Section I, we have presented a method for estimating current popularity to enable artists to plan for improving their popularity. However, prediction of future popularity is also useful; therefore, we should predict future popularity with consideration of time-series data in our future work.

As described in Section IV-C, we should introduce a framework to control the effects of social metadata and multi-modal features in our future work. Also, we will introduce a framework to improve the discriminant power of multi-modal features or select optimal features.

Finally, in the experiment, we evaluated the proposed method in a situation in which there is information about artists, *i.e.*, social metadata and multi-modal features. In other words, we did not evaluate the proposed method in a situation in which there is no information at all about artists. This is because we evaluated whether the proposed method is useful for artists who have recently started to use MSS and for artists who are planning to release new audio tracks or update their biographies. However, there is a great demand for a method that can be used in a situation in which the artist does not have information about "related artists" or does not sufficiently have audio tracks or his/her biography. Therefore, in our future work, we should realize popularity estimation in such situations by introducing a framework to complement links based on existing information about artists.

VI. CONCLUSION

In this paper, we have presented a method for estimating the popularity of artists based on context-aware network analysis. The technical novelty of this paper is construction of a novel framework that can use uni-modal or multi-modal features commonly via analysis of the context-aware network. The technical contribution of this paper is to improve extensibility for using multi-modal features, which enables accurate analysis of latent relationships between artists. The results of experiments using multiple real-world datasets in Spotify confirmed the effectiveness of the proposed method.

REFERENCES

[1] IFPI. (2018). *Global Music Report 2018: State of the Industry*. Accessed: Jun. 1, 2018. [Online]. Available: http://www.ifpi.org/downloads/GMR2018.pdf

[2] IFPI. (2017). *Music Consumer Insight Report 2017: Connecting With Music*. Accessed: Jun. 1, 2018. [Online]. Available: http://www.ifpi.org/downloads/Music-Consumer-Insight-Report-2017.pdf

[3] M. Lee, H. Choi, D. Cho, and H. Lee, "Cannibalizing or complementing? The impact of online streaming services on music record sales," *Procedia Comput. Sci.*, vol. 91, pp. 662–671, 2016.

[4] O. Celma and P. Cano, "From hits to niches? Or how popular artists can bias music recommendation and discovery," in *Proc. ACM KDD Workshop Large-Scale Recommender Syst. Netflix Prize Competition*, 2008, pp. 1–8.

[5] A. Bellogın, A. de Vries, and J. He, "Artist popularity: Do Web and social music services agree," in *Proc. AAAI Conf. Web Social Media*, 2013, pp. 673–676.

[6] M. Schedl, "Analyzing the potential of microblogs for spatio-temporal popularity estimation of music artists," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 539–553.

[7] M. Schedl, T. Pohle, N. Koenigstein, and P. Knees, "What's Hot? Estimating country-specific artist popularity," in *Proc. Int. Soc. Music Inf. Retr. Conf.*, 2010, pp. 117–122.

[8] T. Trzcinski and P. Rokita, "Predicting popularity of online videos using support vector regression," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2561–2570, Nov. 2017.

[9] F. Gelli, T. Uricchio, M. Bertini, A. Del Bimbo, and S.-F. Chang, "Image popularity prediction in social media using sentiment and context features," in *Proc. 23rd ACM Int. Conf. Multimedia - MM*, 2015, pp. 907–910.

[10] X. Huang, Y. Gao, Q. Fang, J. Sang, and C. Xu, "Towards SMP challenge: Stacking of diverse models for social image popularity prediction," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 1895–1900.

[11] W. Wang and W. Zhang, "Combining multiple features for image popularity prediction in social media," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 1901–1905.

[12] F. Figueiredo, "On the prediction of popularity of trends and hits for user generated videos," in *Proc. 6th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2013, pp. 741–746.

[13] J. Lv, W. Liu, M. Zhang, H. Gong, B. Wu, and H. Ma, "Multi-feature fusion for predicting social media popularity," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 1883–1888.

[14] Y. Mao, Y. Shen, G. Qin, and L. Cai, "Predicting the popularity of online videos via deep neural networks," 2017, *arXiv:1711.10718*. [Online]. Available: http://arxiv.org/abs/1711.10718

[15] G. Fontanini, M. Bertini, and A. Del Bimbo, "Web video popularity prediction using sentiment and content visual features," in *Proc. ACM Int. Conf. Multimedia Retr. (ICMR)*, 2016, pp. 289–292.

[16] L. Li, R. Situ, J. Gao, Z. Yang, and W. Liu, "A hybrid model combining convolutional neural network with XGBoost for predicting social media popularity," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 1912–1917.

[17] M. Meghawat, S. Yadav, D. Mahata, Y. Yin, R. Ratn Shah, and R. Zimmermann, "A multimodal approach to predict social media popularity," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, pp. 190–195.

[18] L.-C. Yu, Y.-H. Yang, Y.-N. Hung, and Y.-A. Chen, "Hit song prediction for pop music by siamese CNN with ranking loss," 2017, *arXiv:1710.10814*. [Online]. Available: http://arxiv.org/abs/1710.10814

[19] C.-C. Hsu, Y.-C. Lee, P.-E. Lu, S.-S. Lu, H.-T. Lai, C.-C. Huang, C. Wang, Y.-J. Lin, and W.-T. Su, "Social media prediction based on residual learning and random forest," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 1865–1870.

[20] B. Wu, W.-H. Cheng, Y. Zhang, Q. Huang, J. Li, and T. Mei, "Sequential prediction of social media popularity with deep temporal context networks," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3062–3068.

[21] T. Trzciński, P. Andruszkiewicz, T. Bocheński, and P. Rokita, "Recurrent neural networks for online video popularity prediction," in *Proc. Int. Symp. Methodol. Intell. Syst.*, 2017, pp. 146–153.

[22] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, Aug. 2010.

[23] X. He, M. Gao, M.-Y. Kan, Y. Liu, and K. Sugiyama, "Predicting the popularity of Web 2.0 items based on user comments," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2014, pp. 233–242.

[24] X. Niu, L. Li, T. Mei, J. Shen, and K. Xu, "Predicting image popularity in an incomplete social media community by a weighted bi-partite graph," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 735–740.

[25] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in Twitter," in *Proc. 20th Int. Conf. Companion World Wide Web (WWW)*, 2011, pp. 57–58.

[26] S. Yu and S. Kak, "A survey of prediction using social media," 2012, *arXiv:1203.1647*. [Online]. Available: http://arxiv.org/abs/1203.1647

[27] I. Pitas, *Graph-Based Social Media Analysis*. Boca Raton, FL, USA: CRC Press, 2016.

[28] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, p. 321, Dec. 1936.

[29] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 855–864.

[30] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

[31] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[32] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[33] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[35] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, nos. 1–7, pp. 107–117, Apr. 1998.

[36] J. Cao, Y. Zhang, R. Ji, F. Xie, and Y. Su, "Web video topics discovery and structuralization with social network," *Neurocomputing*, vol. 172, pp. 53–63, Jan. 2016.

[37] R. Harakawa, T. Ogawa, and M. Haseyama, "Extracting hierarchical structure of Web video groups based on sentiment-aware signed network analysis," *IEEE Access*, vol. 5, pp. 16963–16973, 2017.

[38] A. Narayanan, E. Shi, and B. I. P. Rubinstein, "Link prediction by de-anonymization: How we won the kaggle social network challenge," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2011, pp. 1825–1834.

[39] D. Takehara, R. Harakawa, T. Ogawa, and M. Haseyama, "Extracting hierarchical structure of content groups from different social media platforms using multiple social metadata," *Multimedia Tools Appl.*, vol. 76, no. 19, pp. 20249–20272, May 2017.

[40] Y. Matsumoto, R. Harakawa, T. Ogawa, and M. Haseyama, "Music video recommendation based on link prediction considering local and global structures of a network," *IEEE Access*, vol. 7, pp. 104155–104167, 2019.

[41] L. Backstrom and J. Leskovec, "Supervised random walks: Predicting and recommending links in social networks," in *Proc. ACM Int. Conf. Web Search Data Mining*, vol. 2011, pp. 635–644.

[42] S. Volokhin and E. Agichtein, "Towards intent-aware contextual music recommendation: Initial experiments," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2018, pp. 1045–1048.

[43] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[44] N. Cowan, "The magical number 4 in short-term memory: A reconsideration of mental storage capacity," *Behav. Brain Sci.*, vol. 24, no. 1, pp. 87–114, Oct. 2001.

[45] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27-1–27-27, 2011.

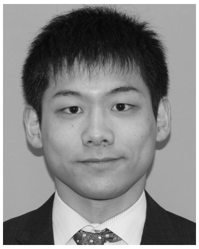
[46] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 701–710.

[47] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. (2003). *A Practical Guide to Support Vector Classification*. Accessed: Jun. 1, 2018. [Online]. Available: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

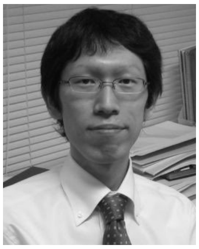
[48] M. Jacomy, S. Heymann, T. Venturini, and M. Bastian, "ForceAtlas2, a continuous graph layout algorithm for handy network visualization," *PLoS ONE*, vol. 9, no. 6, 2014, Art. no. e98679.

[49] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," in *Proc. Int. AAAI Conf. Weblogs Social Media*, 2009, pp. 1–2.

YUI MATSUMOTO (Student Member, IEEE) received the B.S. degree in electronics and information engineering from Hokkaido University, Japan, in 2018. She is currently pursuing the M.S. degree with the Graduate School of Information Science and Technology, Hokkaido University. She also majors in music to receive a B.Mus. degree. Her research interests include music information retrieval and web mining. She is a Student Member of the IEICE.



RYOSUKE HARAKAWA (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Hokkaido University, Japan, in 2013, 2015, and 2016, respectively, all in electronics and information engineering. He is currently an Assistant Professor with the Department of Electrical, Electronics, and Information Engineering, Nagaoka University of Technology. His research interests include multimedia information retrieval and web mining. He is a member of the IEICE and the Institute of Image Information and Television Engineers (ITE).



TAKAHIRO OGAWA (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from Hokkaido University, Japan, in 2003, 2005, and 2007, respectively. He joined the Graduate School of Information Science and Technology, Hokkaido University, in 2008. He is currently an Associate Professor with the Faculty of Information Science and Technology, Hokkaido University. His research interests include AI, the IoT, and big data analysis for multimedia signal processing and its applications. He was the Special Session Chair of the IEEE ISCE 2009, the Doctoral Symposium Chair of ACM ICMR 2018, the Organized Session Chair of the IEEE GCCE, from 2017 to 2019, the TPC Vice Chair of the IEEE GCCE 2018, and the Conference Chair of the IEEE GCCE 2019. He has also been an Associate Editor of *ITE Transactions on Media Technology and Applications*. He is a member of the ACM, IEICE, and ITE.

MIKI HASEYAMA (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics from Hokkaido University, Japan, in 1986, 1988, and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University, as an Associate Professor, in 1994. She was a Visiting Associate Professor at Washington University, USA, from 1995 to 1996. She is currently a Professor with the Faculty of Information Science and Technology, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She has been a Vice-President of the Institute of Image Information and Television Engineers, Japan (ITE), an Editor-in-Chief of *ITE Transactions on Media Technology and Applications*, and the Director of international coordination and publicity with The Institute of Electronics, Information, and Communication Engineers (IEICE). She is a Fellow of the ITE and a member of the IEICE and ASJ.