

ガイド付き LDA を用いた歌詞の分類と可視化

中井 祐希[†] 伊藤 貴之[†]

[†]お茶の水女子大学理学部 〒112-8610 東京都文京区大塚 2-1-1

E-mail: [†]{g1820524, itot}@is.ocha.ac.jp

あらまし 歌詞は重要な音楽の構成要素の1つであり、日常の音楽聴取において大きな影響を与える。そのため、歌詞の意味に基づいた楽曲の探索システムが有用となる。しかし、歌詞に対して抱く印象は主観的であり、歌詞のみに着目して楽曲を分析するのは難しい。そこで本報告では、ガイド付き LDA を用いて歌詞を分析し、歌詞の分布を可視化する手法を提案する。

キーワード 可視化, 歌詞, LDA, 次元削減

1. はじめに

歌詞は重要な音楽の構成要素の1つである[1]。J-POP などの歌唱曲の鑑賞において歌詞が与える影響は大きい。そこで歌詞にもとづいた楽曲の分類・探索が有用となる。しかし歌詞に対して抱く印象は主観的であり、歌詞以外の音楽要素が印象を与える可能性もあるため、ユーザが求める歌詞の探索基準にも個人差が生じやすい。この問題に対して我々は、歌詞の分布を可視化することで、能動的な歌詞探索を支援する研究に取り組んでいる。しかし、論文や記事などと比べて歌詞は語彙の自由度が高いため、満足度の高い可視化結果を得ることが難しい。

そこで本報告では、潜在的ディレクトリ配分法 (Latent Dirichlet Allocation: LDA) を拡張したガイド付き LDA を用いて対話的にガイド単語を入力し、そこから算出された歌詞の分布を可視化する手法を提案する。この手法により、ユーザの視点にもとづいた歌詞の分類結果を、反復的な可視化によって示すことが容易になる。また、他の音楽要素を加味せずに歌詞のみに着目して楽曲群を探索することが可能になる。ユーザは可視化結果を用いることで、楽曲やアーティストの個性や傾向の違い、歌詞の多様さを観察できる。

可視化結果を活用するユーザの対象として、音楽鑑賞者、音楽業界勤務者、作詞者を想定している。音楽鑑賞者は、自分の好きな曲を出発点として、その曲に歌詞が似ている楽曲を可視化結果から探索したり、歌詞の内容から楽曲を探索したりするときに活用できる。また音楽業界勤務者は、歌詞に沿ってアーティストの個性を発見したり、歌詞の流行を俯瞰したりするときに活用できる。さらに作詞者は、各アーティストらしい作詞方法を模索して参考にする場合にも活用できる。

本報告で採用した次元削減と散布図による可視化は、データ全体の分布を俯瞰するのに向いている。この特徴は主に、例えば音楽業界勤務者が「日本のヒット曲全体の傾向を俯瞰したい」というときに向いてい

る。一方で、一般の音楽鑑賞者が音楽を探索する際には、特定のアーティストの楽曲に似た曲を検索したい、特定の内容を歌詞に含む楽曲を探索したい、といった局所的な探索を望むことが多いと考えられる。このような場合には、可視化結果を俯瞰した上で、ユーザが興味をもつ特定の部分にズームする形で楽曲を探索することができる。

本報告の構成は以下のとおりである。2 章では関連研究について述べる。3 章では提案手法について、4 章では本手法の実行例を述べる。5 章では本研究のまとめと今後の課題を述べる。

2. 関連研究

小林[2]らは、歌詞に使用されている単語について、品詞や語種などの語彙指標の使用率の使用率を比較し、言語使用の時系列変化を分析した。

大出[3]らは、単語をグループごとに分け、それぞれのグループにコードを登録し、歌詞に頻出する単語と該当コードを分析することで、歌詞の年代ごとの傾向や文化を探索した。また、定村[4]らは、松本隆が作詞した楽曲の歌詞について計量テキスト分析を用い、歌手別の使用頻度を比較した。

河村[5]は、検索単語の連想語を自動抽出し、TF-IDF 法を用いて歌詞の特徴量を算出することで、検索単語とその連想語によって楽曲推薦をする手法を提案した。また細谷[6]らは、複数の女性シンガーソングライターの歌詞を、ランダムフォレストを用いて探索的に分析した。

Hossain[7]らは、LDA を用いて大量の楽曲の歌詞を分析し、曲名に入れるべき単語を算出することで、歌詞に基づいた曲名を推薦する手法を提案した。また佐々木[8]らは、歌詞の潜在的意味を LDA で求め、多数の既存の歌詞の中から、ユーザが好む歌詞をインタラクティブに検索できる歌詞検索インタフェース

「LyricsRadar」を提案した。

本手法では、歌詞群の分布算出にガイド付き LDA を用いる点で既存研究と異なる。また、ガイドごとに散布図を複数表示させることで、楽曲の分布や類似性に対して多様な説明が可能になるという点でも既存研究とは異なる。

3. 処理手順

本研究の処理手順を示す。本研究ではまず楽曲データを収集し、この各曲の歌詞にガイド付き LDA を適用することで、歌詞を数値化（ベクトル化）する。このベクトルに次元削減を適用し、歌詞の分布を散布図として可視化する。本研究では実行結果の比較のために、ガイド付きではない一般的な LDA を適用した結果も可視化する。

3.1. 楽曲データの収集

本研究では、タイトル、アーティスト、作詞家、年代、歌詞で構成される楽曲データを構築した。現時点での本研究では、1988 年から 2007 年の CD シングルの売り上げが高い上位 10 曲と、2008 年から 2020 年の Billboard Japan の年間チャート上位 10 曲を題材として楽曲データを収集した。その結果として、現時点では 328 曲分の楽曲データを採用している。

歌詞の形態素解析には MeCab を使用した。名詞、動詞、形容詞、副詞、感動詞、連体詞のみを抽出し、その原型を 1 単語として数えた。ただし、極端に多くの楽曲に頻繁に出現する単語は、複数のトピックの重要語として選出されるため、ストップワードに設定し、分析対象から外した。

3.2. 歌詞の分析

3.2.1. LDA を用いた歌詞の分析

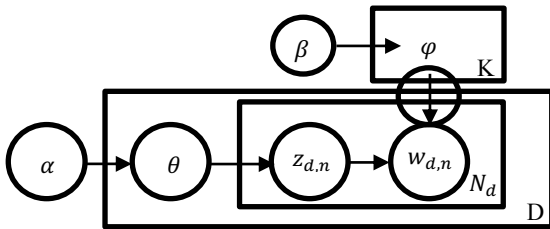


図 1 LDA のグラフィカルモデル

潜在的ディレクトリ配分法 (Latent Dirichlet Allocation: LDA)[9]は、各文書が複数のトピックから構成されているとするモデルである。図 1 は LDA のグラフィカルモデルを表している。顕在変数は網掛けの頂点、潜在変数と未知パラメータはその他の頂点、それ

らの依存関係は有向辺で表現される。矩形は、角に記された数だけ、矩形内の変数の生成が繰り返されることを示している。本手法では 1 つの歌詞を 1 つの文書とし、 D 個の独立した歌詞 $X = \{X_1, \dots, X_D\}$ を考える。歌詞 X_d は N_d 個の単語 $X_d = \{w_{d,1}, \dots, w_{d,N_d}\}$ で構成されており、各単語は背景にトピックを持つとする。 K はトピック数、 θ は各文書のトピック多項分布パラメータ、 ϕ は各トピックの単語多項分布パラメータ、 α と β はそれぞれ θ と ϕ のディリクレハイパーパラメータである。グラフィカルモデルから文書生成過程を以下に示す。

- (1) 各文書 X_d に、 $\theta_d \sim \text{Dir}(\alpha)$ を選択する。
- (2) 各トピック k に、 $\phi_k \sim \text{Dir}(\beta)$ を選択する。
- (3) 文書 X_d の N_d 個の各単語 $w_{d,i}$ に対して、
 - トピック $z_{d,i} \sim \text{Mult}(\theta_d)$ を選択する。
 - 単語 $w_{d,i} \sim \text{Mult}(\phi_{z_{d,i}})$ を選択する。

本手法では、トピックモデルの推定にギブスサンプリング[10]を用いる。文書 X_d の i 番目の単語に割り当てられるトピックを除き、全てが定まっていると仮定する場合、そのトピックが k である確率は、以下の式で与えられる。

$$P(z_{d,i} = k | w_{d,i} = w, z_{-i}, w_{-i}, \alpha, \beta) \propto \frac{C_{kd,-i}^{KD} + \alpha}{\sum_{k'} C_{k'd,-i}^{KD} + K\alpha} \frac{C_{wk,-i}^{WK} + \beta}{\sum_{w'} C_{w'k,-i}^{WK} + K\beta}$$

w_{-i} は w から w_i を除外した集合、 z_{-i} は z から z_i を除外した集合である。 $C_{kd,-i}^{KD}$ はトピック k が文書 X_d に割り当てられた回数、 $C_{wk,-i}^{WK}$ は単語 w にトピック k が割り当てられた回数であり、 i 番目の単語は共に除外している。本手法では、各歌詞に対するトピック多項分布パラメータ θ_d (すなわち歌詞のトピック混合比) の期待値を求め、3.3 項で後述するように次元削減を適用することで、2 次元平面上に楽曲の分布を可視化した。

3.2.2. ガイド付き LDA を用いた歌詞の分析

ガイド付き LDA[11]は、あらかじめ重要な単語を予約語（ガイド単語）として各トピックに割り当てておく LDA である。これにより、各トピックに代表語として分類される単語は、ガイド単語とそれに共起する単語である可能性が高くなる。そのため、トピックの分類結果に明確な視点を導入することが可能になる。ここでは LDA と同様に、各歌詞に対するトピック混合比の期待値を求める。本報告では、トピックテーマとガイド単語を以下の 2 種類

- (1) 季節ごとに設定したトピックテーマ
 - (2) イベントごとに設定したトピックテーマ
- に沿って設定した結果を示す。本報告で採用しているトピックテーマと予約語を表 1 に示す。

表1 ガイド付き LDA におけるトピックテーマと予約語

(1) 季節ごとの場合	
トピックテーマ	ガイド語
春	春, 桜, 卒業
夏	夏, 祭り, 花火, ひまわり
秋	秋
冬	冬, 雪

(2) イベントごとの場合	
トピックテーマ	ガイド語
卒業	卒業, 桜, 制服, 友, 別れ, 門出, サヨナラ
夏祭り	夏, 祭り, 花火, 氷菓子

3.2.3. 学習時における単語の扱い

本報告ではガイド付き LDA による学習に加えて, 比較のためにガイド付きでない通常の LDA による学習結果も求めた. ここで, 日本語詞と英語詞が混合している歌詞は, 日本語詞の部分のみを LDA の分析の対象とした.

本報告では前処理として, ストップワードを設定せずに LDA を全楽曲に適用してみた. その結果, 複数のトピックに対して同時に重要語とされた単語が見られた. この単語のうち, トピックの形成において我々の主観で重要でないと考えた以下の単語群(表2参照)を, ストップワードとして設定した.

表2 ストップワード

する, ある, ない, いる, なる, この, その, あの, それ, 僕, あなた, 君, 私, 僕ら, 俺

3.3. 可視化

LDA またはガイド付き LDA を用いて算出した各歌詞の各トピックへの期待値をベクトルとみなして, t-SNE[12]を用いてこれを2次元に次元削減し, 散布図として可視化した. この散布図において, 各々の点は各々の歌詞に対応する.

4. 実行例

本手法の実行例を示す. 本報告ではトピック数を $K=4$ として学習した.

4.1. トピックテーマ: 季節

歌詞に「夏」という単語を含む歌詞を赤色の点, その他の歌詞を青色の点として散布図に表示した結果を図2に示す. LDA を用いた場合と比べて, ガイド付き LDA を用いた場合の方が独立した点が少なく, 歌詞の類似度に沿って歌詞を配置することができた. 散布図中の3個以上点が集まっている部分(図2の青丸参照)

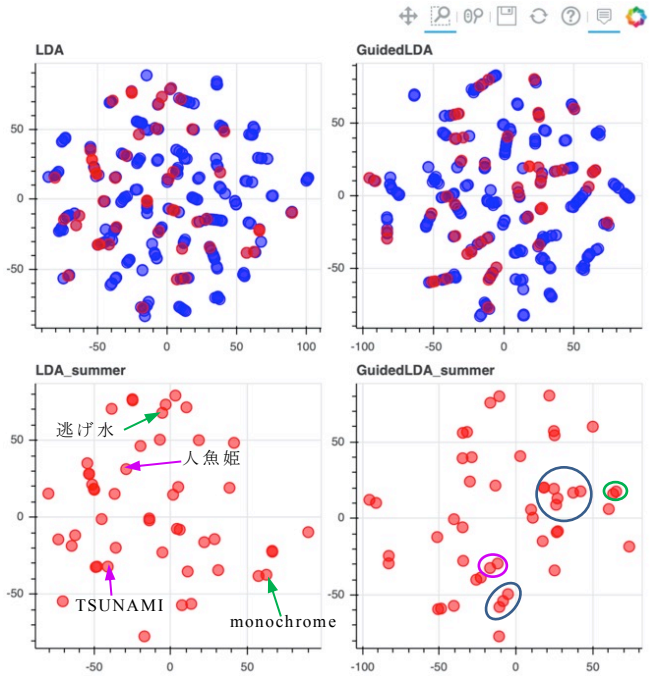


図2 (左上)LDA の適用. (右上)ガイド付き LDA の適用. (左下)「夏」を含む歌詞のみ(LDA). (右下)「夏」を含む歌詞のみ(ガイド付き LDA).

に注目すると, 全て夏に関係があり, 恋愛をテーマにした歌詞であった. ガイド付き LDA を適用した場合の散布図で重なっている2つの点のうち, LDA を適用した場合は点と点の距離が離れていたものに注目する. 図2にて緑丸に含まれる2つの点は, 乃木坂46の「逃げ水」と浜崎あゆみの「monochrome」である. この2曲は, 失恋をテーマとしており, 夏の恋を「夢」や「幻」という単語で例えている点が共通している. また, 図2にて紫丸に含まれる2つの点は, 中山美穂の「人魚姫」と桑田佳祐の「TSUNAMI」である. この2曲は過去の夏の恋に関係する歌詞であり, どちらも歌詞の中に「雨」という単語が多用されている. このように, 単に夏に関係するだけでなく情景描写にも共通点がある, という楽曲を近くに配置できている点でガイド付き LDA のほうが良好な配置結果をもたらしている.

続いて, 「夏」と「花火」を1つのトピックのガイド語に設定し, それ以外のトピックにガイド語を設けなかった場合の可視化結果を示す(図3参照). 歌詞に「夏」という単語を含む歌詞を赤色の点, その他の歌詞を青色の点で表示した. また, Mr. Children の「HANABI」と, たまの「さよなら人類」の2曲を緑色の点で表示した. この2曲は歌詞に「花火」という単語を含むが, 歌詞全体は夏に関係しない曲である(表3参照).

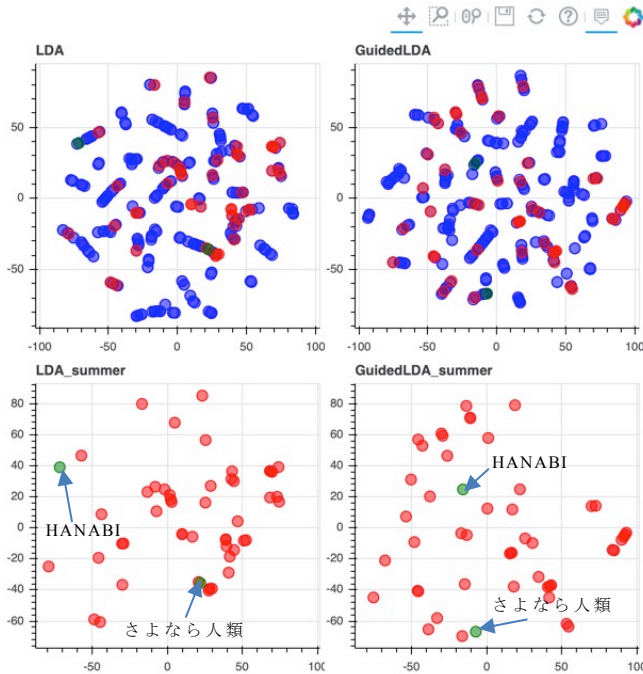


図3 (左上)LDAの適用. (右上)ガイド付き LDA の適用. (左下)「夏」を含む歌詞と「HANABI」,「さよなら人類」のみ(LDA). (右下)「夏」を含む歌詞と「HANABI」,「さよなら人類」のみ(ガイド付き LDA).

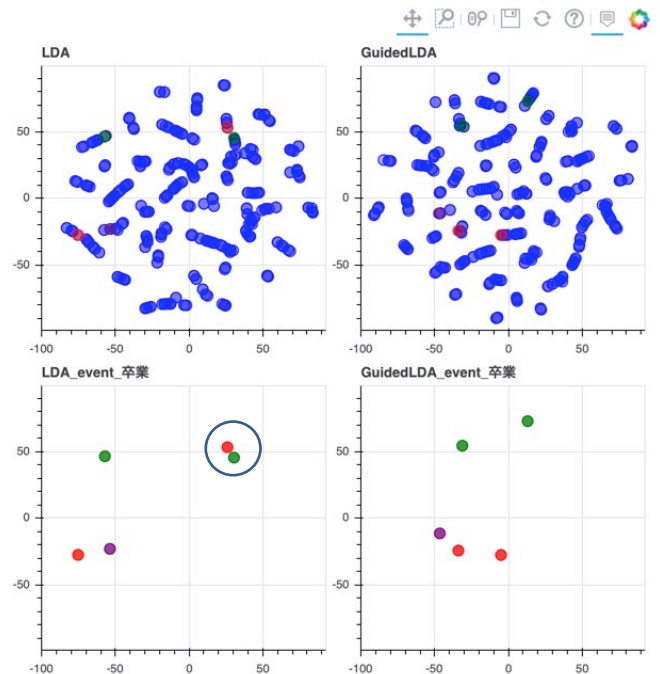


図4 (左上)LDAの適用. (右上)ガイド付き LDA の適用. (左下)「卒業」を含む歌詞と卒業に関する内容の歌詞のみ(LDA). (右下)「卒業」を含む歌詞と卒業に関する内容の歌詞のみ (ガイド付き LDA).

表3 歌詞中で登場する「花火」の例

楽曲名	具体的な歌詞
HANABI	決して捕まえることの出来ない 花火のような光だとしたって
さよなら人類	あのこは花火を打ち上げて この日が来たのを祝ってる

「HANABI」は, LDA を用いた場合とガイド付き LDA を用いた場合のどちらの散布図上でも, 独立した点である. しかし, ガイド付き LDA を用いた場合の方が, LDA を用いた場合と比べて, 近くの点との距離が離れている. また, 「さよなら人類」は, LDA を用いた場合の散布図上では, ORANGE RANGE の「ロコローション」などの夏に関連する曲が集まっている部分に点が配置されている. しかし, ガイド付き LDA を用いた場合の散布図上では独立した点となっている. 言い換えれば, ガイド付き LDA のほうが, 「夏」にだけ関係ある楽曲, 「花火」にだけ関係ある楽曲をうまく分離した配置をもたらしていると言える.

4.2. トピックテーマ: イベント

歌詞に「卒業」という単語を含み, 卒業に関する内容の歌詞を赤色の点, 「卒業」という単語は含んでいないが, 卒業に関する内容の歌詞を紫色の点, 「卒業」という単語を含んでいるが, 歌詞全体は卒業に関係しない歌詞を緑色の点, その他の歌詞を青色の点として, 散布図上に表示した結果を示す(図4参照). LDA を用いた場合の散布図上では, 歌詞に「卒業」という単語を含み, 卒業に関連する歌詞(赤色の点)と「卒業」という単語を含むが, 卒業に関連しない歌詞(緑色の点)の距離が近くなっている(青丸参照). 一方でガイド付き LDA を用いた場合の散布図では, 「卒業」という単語を含み, 卒業に関連する歌詞は, 「卒業」という単語を含むが, 卒業に関連しない歌詞と, 点と点の距離が離れている. また, 卒業に関連する歌詞(赤色の点と紫色の点)は, 「卒業」という単語の有無に関わらず, 点と点の距離が近くに配置されている. 以上により, ガイド付き LDA のほうが, 卒業に関連する楽曲と関連しない楽曲をうまく分離した配置をもたらしているといえる.

また, 歌詞に「祭り」という単語を含み, 夏祭りに関連する内容の歌詞を赤色の点, 「祭り」という単語を含むが, 夏祭りに関連しない内容の歌詞を緑色の点と

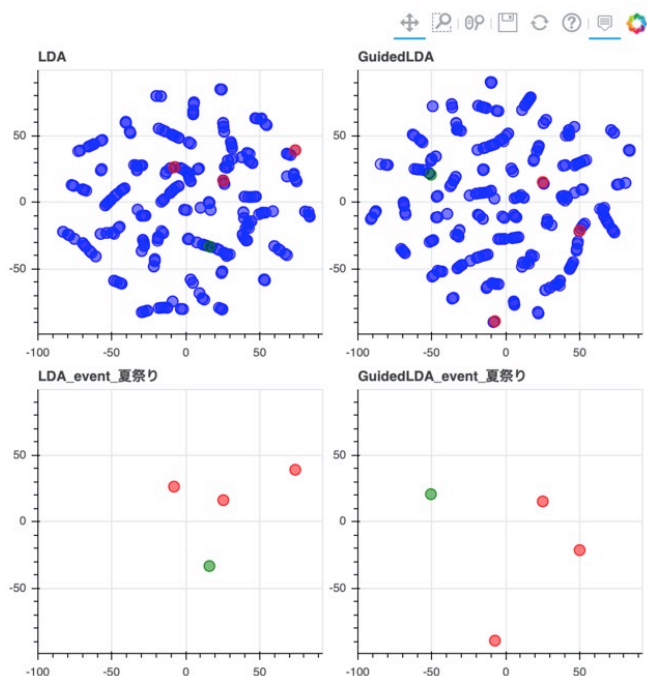


図5 (左上)LDAの適用。(右上)ガイド付きLDAの適用。(左下)「祭り」を含み夏祭りに関する歌詞と「祭り」を含むが夏祭りに関係しない内容の歌詞のみ(LDA)。(右下)「祭り」を含み夏祭りに関する歌詞と「祭り」を含むが夏祭りに関係しない内容の歌詞のみ(ガイド付きLDA)。

して、散布図上に表示した結果を示す(図5参照)。LDAを用いた場合とガイド付きLDAを用いた場合のどちらの散布図上でも、夏祭りに関連する歌詞と夏祭りに関連しない楽曲は離れて配置されている。しかし、ガイド付きLDAを用いた場合の方が、LDAを用いた場合と比べて、点と点の距離がさらに離れていることがわかる。

5. まとめと今後の課題

本報告では、歌詞の分布を可視化する手法として、ガイド付きLDAを用いて歌詞のトピックを数値化し、これに次元削減を適用して散布図として表示する手法を提案した。ガイド語を選択することで、歌詞の分布の可視化結果にユーザ個人の視点が導入され、歌詞の類似性や相違性をユーザ個人の視点から解釈することが容易になる。また、ガイド語の選択によって多様な可視化結果を示すことが可能であるため、多様な視点から歌詞の分布を比較することが容易になる。ガイド付きではない一般的なLDAを適用した場合と比較すると、ガイド付きLDAを適用した場合のほうが、情景描写の共通性や全体的な歌詞の内容を考慮した分布を算出することができた。

今後の課題は以下の通りである。

まず1点目として、ガイド語の選出方法の模索があげられる。著者が実験した限りでは、設定したガイド語によっては歌詞の内容を反映した可視化結果が得られない場合があった。どのようなガイド語を設定することでユーザが納得できる可視化結果が得られるかについて、さらに実験を重ねたい。

2点目は英語詞を中心とした歌詞への対応である。現在は英語詞を省略して日本語詞の部分だけを処理の対象としている。英語詞も対象として処理をしたら、これを日本語詞に和訳するのか、それとも英単語は和単語とは別の単語として扱うのか、議論が分かれるところである。この点について議論を深めたい。

3点目は可視化の色表現である。各楽曲に対応する散布図上の各点の色を、アーティストや作詞家ごとに割り当てたり、年代ごとに割り当てたりすることで、アーティストの個性や年代ごとの歌詞の傾向を俯瞰することが可能である。多様な色表現を実装することで、歌詞の内容とメタ情報の多様な相関を発見したい。

4点目はユーザインタフェース機能の実装である。本研究では「ガイド単語を入力して可視化結果を確認するという一連の操作を反復することで、ユーザの目的に合った可視化結果を得る」ことを目標としているが、それを実現するための対話操作機能をまだ実装していない。ガイド単語を入力すると歌詞の分布の可視化結果が対話的に表示される機能を実装して、その効果を検証したい。

以上について検討したのちに、さらに多くの楽曲数で可視化結果を検証し、ユーザ評価実験を実施したい。

参考文献

- [1] 森一馬, “日常の音楽聴取における歌詞の役割についての研究”, 対人社会心理学研究, 10, 31-137, 2010.
- [2] 小林雄一郎, 天笠美咲, 鈴木崇史, “語彙指標を用いた流行歌の歌詞の通時的分析”, じんもんこん 2015 論文集, 23-30, 2015.
- [3] 大出彩, 松本文子, 金子貴昭, “流行歌から見る歌詞の年代別変化”, じんもんこん 2013 論文集, 4, 103-110, 2013.
- [4] 定村薫, “松本隆の歌詞の使用単語についての計量テキスト分析”, 尚美学園大学総合政策研究紀要, 34, 17-33, 2019.
- [5] 河村康治, “歌詞情報の歌詞情報の分析に基づくユーザの状況を考慮した楽曲推薦に関する研究”, 大学院研究年報, 理工学研究科, 47, 2017.
- [6] 細谷舞, 鈴木崇史, “女性シンガーソングライター の歌詞の探索的分析”, じんもんこん 2010 論文集, 15, 195-202, 2010.
- [7] Hossain, R., Sarker, M.R.K.R., Mimo, M., Al Marouf, A., Pandey, B., “Recommendation Approach of English Songs Title based on Latent Dirichlet

Allocation applied on Lyrics”, 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 1-4, 2019.

- [8] 佐々木将人, 吉井和佳, 中野倫靖, 後藤真孝, 森島繁生, “ LyricsRadar: 歌詞の潜在的意味に基づく歌詞検索インタフェース”, 情報処理学会論文誌, 57.5, 1365-1374, 2016.
- [9] Blei, David M., Andrew Y. Ng, Michael I. Jordan, "Latent dirichlet allocation", the Journal of machine Learning research, 3, 993-1022, 2003.
- [10] Griffiths, T.L., Steyvers, M., “Finding Scientific Topics”, Proceedings of the National Academy of Sciences of the United States of America, 101, 5228-5235, 2004.
- [11] Jagarlamudi, J., Daumé III, H., Udupa, R., “Incorporating lexical priors into topic models”, Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 204-213, 2012.
- [12] Van der Maaten, Laurens, and Geoffrey Hinton, “Visualizing data using t-SNE”, Journal of machine learning research, 9.11, 2008.