7ZE-07

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 〒112-8610 東京都文京区大塚 2-1-1

1. 概要

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．

2. 関連研究

河村[2]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[3]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[4]らは，LDA を用いて大量の楽曲の歌詞を分析し，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[5]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. 提案手法

3.1. 楽曲データの収集

タイトル，アーティスト，作詞家，年代，歌詞を 1 つの楽曲データとする．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 曲の楽曲データを収集した．ただし，英語詞だけで構成される楽曲は対象から外した．また，複数の年に登場する楽曲は年代が最も新しい楽曲データのみを使用し，楽曲データの重複が起こらないようにした．本報告では，全部で 332 曲分の楽曲データを使用することにした．

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

LDA[6]は，文書群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．本手法では，1 つの歌詞を 1 つの独立した文書とし，各歌詞に含まれる単語はそれぞれ，背景にトピックを持つとする．また，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．各歌詞に対するトピック混合比の期待値を求め，次元削減することで楽曲を二次元平面状に配置した(3.3 節で後述)．

3.2.2. ガイド付き LDA を用いた歌詞の分析

ガイド付き LDA[8]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

Classification and visualization of lyric collections using guided LDA
Yuki NAKAI[†]　Takayuki ITOH[†]
†Department of Information Sciences, Ochanomizu University

マとガイド単語を以下のように設定した場合の結果を示す(表 1 参照).

表 1　ガイド付き LDA におけるトピックテーマと予約語

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

3.2.3. LDA の学習

本報告ではトピック数を$K = 4$として学習した．また，ストップワードを設定せずに LDA を用いて全ての歌詞を分析した際に，トピックの重要語として頻出した以下の単語を，ストップワードとして設定した(表 2 参照)．

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

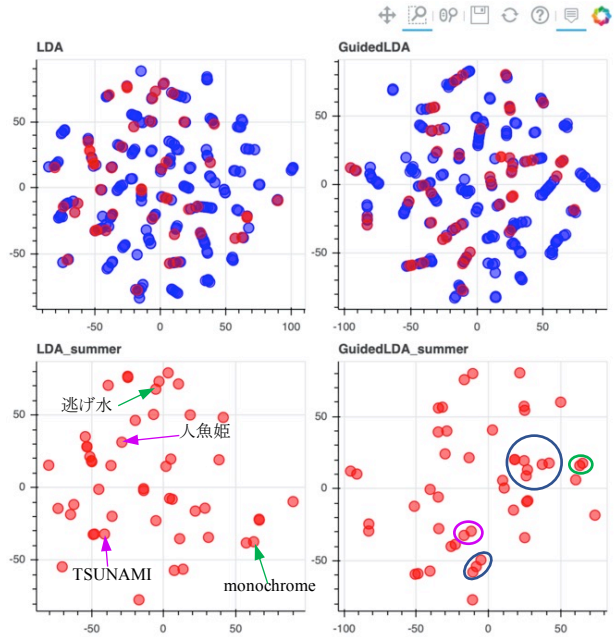

図 2 (左上)LDA の適用．(右上)ガイド付き LDA の適用．(左下)「夏」を含む歌詞のみ(LDA)．(右下)「夏」を含む歌詞のみ(ガイド付き LDA)．

LDA とガイド付き LDA を用いた歌詞の分析結果を，t-SNE を用いて 5 次元から 2 次元に次元削減し，散布図として可視化した．1 つの点が 1 つの歌詞を表す．本報告では，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状に表示した結果を示す(図 2 参照)．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)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．

4.　まとめと今後の課題

本報告では，歌詞を分類し可視化する手法として，ガイド付き LDA を用いて歌詞を分析し，散布図状に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の説明が可能になる．また，選択するガイド語を変えることにより，複数の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．

今後の課題としてまず，英語詞を中心とした歌詞に対応することがあげられる．そのほかの課題としては，さらなる楽曲データの収集や，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ユーザインタフェース機能の実装があげられる．

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]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.
[3]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.
[4] 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.
[5] 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.
[6] Blei, David M., Andrew Y. Ng, Michael I. Jordan, "Latent dirichlet allocation", the Journal of machine Learning research, 3, 993-1022, 2003.
[7] 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.
[8] 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.