3Q-04

「作詞家の歌詞の計量テキスト分析と年代推定」

池澤和希　　浦谷則好
東京工芸大学　工学部　コンピュータ応用学科

１　はじめに

歌は常に人々と共にあり，時に元気づけ,時に心を癒し，いつの時代も人々の関心を集め続けている.人気の歌は話題を作り，人々にとって歌とは生活の一部とも言える存在となっている.

時代と共に流行は変化し，受け入れられる歌もそれに合わせ変化し続けている.

こうした人間と常に寄り添って歩んできた歌を調べることにより社会学的にも有用な人の動向，興味を調べることが可能となる.

そこで本研究では歌を形成する要素である歌詞に焦点をあて，詞の特徴や傾向を分析し，さらに歌詞を実際に作成した作詞家についても分析を行うことにする.

２　関連研究

テキストマイニングの手法を用いた研究は近年増加し続けている.その中でも本研究同様に歌詞に焦点を当てた研究も進められている.細谷舞らの「女性シンガーソングライターの歌詞の探索的分析」[1]は研究対象が近く,成分をプロット上に表示し可視化し,アーティストの歌の傾向を分析している.本研究では作詞家の頻出語や特徴量から見られる分析だけでなく語と語,作詞家と作詞家の繋がりを重視した分析を行う.

３　計量テキスト分析

３．１　分析対象

本研究では作詞家の歌詞にどのような特徴・傾向があるかという分析を行う．演歌，J－POP,ロック等様々な楽曲の歌詞を収集すべく作詞家30 人の歌詞を 1 人あたり 100 曲程度収集した.重複している楽曲を省いたため作詞家毎の対象件数は必ずしも同じでない.

３．２　データ

研究に使用する歌詞データは魔境[2]という歌詞検索サイトから抽出する.

Qantitative Text Analaysis and Generation Estimate for words of songs
Kazuki Ikezawa　Noriyoshi Uratani
Tokyo Polytechnic University

魔境はフリーワード検索で「歌手名」「曲名」「歌詞」「アルバム」の４種類の検索が可能である．従来の歌詞検索サイトと違い歌詞中の語から検索ができる．歌詞中に記載されている作詞家の情報を利用しデータを収集する.

３．３　歌詞テキストファイル

前述の魔境から作詞家での検索を行う．検索結果が最大 100 件表示されるため，その 100 件の曲が表示されているページのソースを抽出し,得られたデータから各楽曲のページへ移動し歌詞を抽出しテキストファイルとして保存する.なお作詞家を入力して検索した場合でも作曲家として歌詞中に表記されているものも含まれてしまうため，それら不要なデータはパターンマッチを行うことで取り除く．さらに得られた歌詞は日本語が簡体字（中国語）で書かれていることが多いため， Kanconvit[3]を用いて日本語漢字へ変換しておく

３．４　分析手法

本研究では立命館大学産業社会学部の樋口耕一氏が開発した KHCoder[4]を分析に用いる.KHCoder ではデータ中から語を自動的に抽出し,結果を集計・解析することで特徴を探り要約,グラフにし可視化することが可能である.

KHCoder を利用して共起ネットワーク分析と対応分析を行う.

３．４．１　共起ネットワーク分析

傾向・特徴の分析にあたって語と語の結びつきを探る共起ネットワーク分析を用いる.収集した歌詞テキストファイルを作詞家ごとに結合し,KHCoder を使用して共起ネットワーク図を生成する.

共起ネットワーク図では円の大きさは単語の出現数を指し,線の太さは共起関係が強いことを指す.円の色は中心性を意味する.単語間の距離は jaccard 係数を用いる.なお,円と円の布置された距離は大きな意味を持たない.

３．４．２　対応分析

テキストの部分ごとの特徴を探るため対応分

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.

析を用いる．共起ネットワーク分析と同様に結合したファイルを使用し，2次元プロットを作成する．対応分析では布置された語が近くにあるほど関連が強いことを示す．作詞家の近くに布置されている語を見ることで各作詞家と関連の強い語が，作詞家の近くに布置されている作詞家を見ることで作詞家同士の特徴・類似性が把握できる．

4　年代推定

年代を推定するため，年代の情報を持つデータを新たに作成する．年代流行[5]という歌のランキングを年代ごとに表示できるサイトから 1950 年代～2000 年代までの楽曲を 10 年区切りで 6 つの年代分のリストを作成する．歌詞の取得には J-Lyric.net[6]を用いて曲名と歌手名をリストから入力しヒットした楽曲の歌詞のみ収集し，歌詞から使われている単語を抽出する．

上位 150 語を用いてベクトル化し，各曲とのコサイン類似度を求め年代推定を行う．

5　結果・考察

共起ネットワーク分析では，それぞれの作詞家毎に分析を行った．使用頻度の高い語や中心的に使う語を可視化することに成功した．

一例として図 1 に藤林聖子氏の分析結果を示す．「夢」や「信じる」が頻出し，「夢」「未来」「心」が中心性の高い語となっている．さらに，「夢」からは「出来る」や「信じる」が強い共起を示していることなどから，前向きな歌詞が多いことが推測される．

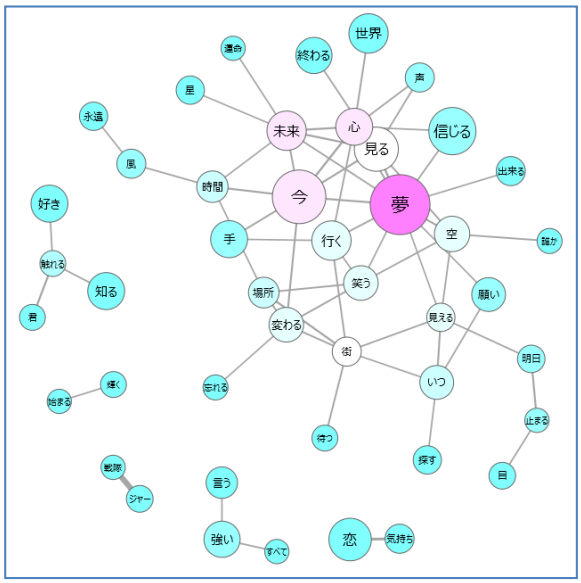

図 1　藤林聖子の共起ネットワーク図

対応分析では，全作詞家を一つにまとめて分析を行った．その結果，図 2 のようにそれぞれの作詞家と関係の強い単語をわかりやすく表示することができた．また，水木れいじやたかたかし，高田ひろおといった主に演歌を作詞することの多かった作詞家が近い所に布置されるなど作詞家同士での似た傾向を持つ人物が明らかになった．

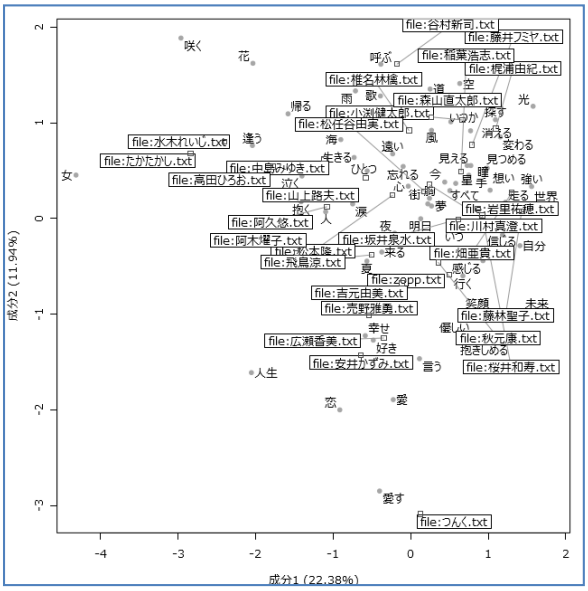

図 2　全作詞家の対応分析表

年代推定では，各曲と年代特徴ベクトルとの類似度を求めたが各歌詞の本来の年代と合致するものは少なかった．1950 年代は「花」，1970 年代は「雨」，2000 年代には「未来」など mass としての特徴は把握できたが，曲単位での分類は出来なかった．

6　おわりに

本研究では，取得した歌詞・作詞家の情報を複数の方法を用いての分析と可視化を行った．さらに，語の傾向から楽曲の年代を推定する手法を提案した．

今後の課題として，年代毎の曲数を増やしたり，作詞家の情報を増やしたりして分析の精度を向上させることが考えられる．さらに，別の可視化手段を用いて単純に見た場合に理解をしやすくする手法の開発が挙げられる．

参考文献

[1]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: 人文科学とコンピュータシンポジウム論文集
[2] http://mojim.com/
[3] http://kanconvit.ta2o.net/
[4] http://khc.sourceforge.net/
[5] http://nendai-ryuukou.com/
[6] http://j-lyric.net/

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.