SNSからファン層は見えるのか？
-Twitterを利用した音楽アーティストのファン特性の抽出-

三村　乃那†　　牛尼　剛聡††

† 九州大学芸術工学部　〒 815–8540 福岡県福岡市南区塩原 4-9-1
†† 九州大学大学除芸術工学研究所　〒 815–8540 福岡県福岡市南区塩原 4-9-1
E-mail: †1DS13214K@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　ソーシャルメディアを利用するユーザの増大に伴い，一般の人々により膨大かつ多様な発言がソーシャルメディア上に投稿されている．ソーシャルメディアには，様々な対象に対する意見が含まれているため，ソーシャルメディアは様々な事物の特徴を知るための情報源としての役割を持つと考えられ，ソーシャルメディアを利用して様々な事物の特性を抽出することが期待されている．本研究では，代表的なマイクロブログのひとつである Twitter におけるユーザの発言の日常性に着目し，特定の音楽アーティスト等に対するフォロワーが投稿する日常的な話題からファン層を推測する手法を提案する．

キーワード　Twitter，ファン層，LDA，トピックモデル

1. はじめに

1.1 背　　景

近年，ソーシャルネットワーキングサービス（SNS）を始めとしたソーシャルメディアの利用者数は増加し続けている [1]．ソーシャルメディア上には，膨大なユーザが存在し，継続的な投稿が行われている．SNS に投稿される記事の対象や内容は多種多様であるが，人々の様々な意見や嗜好が反映されている．そうしたことから，SNS に投稿された記事を分析することで，実世界の様々なモノやコトに対する人々の意見を分析し，それらの社会的な位置づけを抽出できることが期待できる．

これまでにも，SNS に投稿された記事に含まれる特定の商品やコンテンツに対する意見や感想の投稿に着目し，ソーシャルメディアをマーケティングに活用するための手法が提案されてきた．それらの多くは，SNS のユーザにおける日常的な投稿に記載された対象への評価や言及を分析して，対象の特徴を抽出するアプローチである．しかし，対象とする商品やコンテンツを支持する人（ファン）の SNS 上の日常的な投稿内容から，その人々の特徴が抽出できれば，ジャンルが異なる様々なアイテムに対して，そのターゲットとなるユーザの特徴の類似性を評価でき，商品推薦等に応用できる可能性がある．

本研究では，音楽アーティストのファンを対象に，SNS における日常の投稿を分析することで，音楽アーティストごとのファン層を抽出する手法を提案し，評価実験により提案手法の有効性を評価する．

1.2 アプローチ

本研究では，対象とする SNS として，代表的なマイクロブログのひとつである Twitter(注1) を対象とする．そして，Twitter を利用して音楽アーティストのファン特性の抽出を行うことを目的とする．Twitter における投稿の単位はツイートと呼ばれ，一般的にユーザの興味や身近で起こった出来事などのユーザの日常に関する文章が投稿されることが多い．多くの音楽アーティストが Twitter 上にアカウントを持ち，日常的に頻繁に投稿を行なっている．また，音楽アーティストのファンは，自分が好きな音楽アーティストをフォローすることが多く行われている．本研究では，音楽アーティストのフォロワーに対して，日常的なツイートを分析し，中心的なファン層を抽出し，音楽アーティスト自身を特徴づける．

Twitter のフォロワーの日常的なツイートから，その特徴を抽出するためには様々なアプローチが考えられる．本研究では，Twitter では一般的にユーザの興味や身近な出来事などのユーザの日常に関する文章が投稿される点に着目し，ユーザの日常的な投稿に含まれる話題を利用する．

ユーザの日常的な話題を利用するためには，対象となる話題を決定しておく必要がある．そこで，本研究では，日常的な話題を多く含んでいる文書集合から LDA を用いてトピックを抽出する．トピックを抽出するための文書集合として，以下の 3 種類の文書集合を利用する．

（1） 生活情報記事サイト All About(注2) の記事
（2） ランダムに取得したユーザのツイート
（3） Wikipedia(注3) の記事集合

上記のそれぞれの文書集合を対象に LDA を利用してトピックモデルを構築し，基底トピックを抽出する．そして，抽出された基底トピックに基づいて，ユーザのツイートにおける基底トピックの分布を推定し，ユーザの特徴とし，ファン集合内における中心的なファンをグループ化して，ファン層を推定する．そして，ファン層のトピックの分布の類似性に基づき，アーティ

(注1)：http://twitter.com
(注2)：https://allabout.co.jp
(注3)：https://ja.wikipedia.org

スト間の類似性を推定する．

2. 関連研究

これまでに，Twitter のデータを利用した情報抽出や情報要約に関する様々な研究が行われてきた．それらの研究の中でも，有名人や商品などのアイテムに対する評価情報を分析し，マーケティングやサービスに応用しようとする研究は数多く存在する [2] [3].

一方で，評価や言及の対象ではなく，ユーザの自身の特徴を抽出する研究も行われており，ソーシャルメディアにおけるユーザの特徴となる情報について議論されている．長浜らは，Twitter 上のユーザの口語体で投稿される文章の構成に着目し，単語ベース，品詞割合ベース，品詞並びベースの 3 つのアルゴリズムを用いた性別判定の手法を提案している [5]．古賀らは，Twitter における対象ユーザの興味や嗜好に着目し，フォロー関係やリツイートなどを文書として LDA を行うことで潜在的なトピックからユーザ間の類似度をはかり，推薦を行っている [7].

本研究では，ユーザの特徴を抽出するために，ユーザ自身が意識して与えているプロフィールやフォロー関係ではなく，日常的なツイートに含まれる潜在的なトピックに着目する．ユーザの属性抽出を行った上で，アイテムに対するファンをグループ化してファン層の属性とし，アイテムの属性を推定するという点で違いがある．

3. 提案手法

本研究では，Twitter におけるファンの投稿における日常的な話題に基づいた音楽アーティストのファン特性を抽出することを目的とする．そのために，話題を抽出して，ユーザが日常的にツイートしている話題の分布を特徴としたファン層を推定する．本手法は，具体的には以下の処理を行う．

（1） ツイートの収集とクレンジング
（2） 基底トピックの分布によるユーザの特徴量算出
（3） 中心的なユーザの抽出

3.1 ツイートデータの収集とクレンジング

音楽アーティストに関するファンの特徴を抽出するための情報源として，ファンが投稿したツイートを利用する．また，話題抽出のためにランダムなユーザのツイートを利用する．なお，ツイートを収集するユーザは，使用言語が日本語であるユーザに限定する．

ユーザは自身の興味や周囲の出来事などの話題を個人的に発信するだけではなく，ツイートの中には他ユーザへのメッセージであるリプライや，他ユーザのツイートの拡散に用いるリツイートが存在する．リプライおよびリツイートは内容が他ユーザに影響されるため，必ずしも本人の興味のある話題とは限らない．他ユーザに影響されたツイートが多く含まれると，ユーザが発信する話題に関する情報の抽出が妨げられる．そのため，本研究では，リプライおよびリツイートは分析の対象から除外する．

Twitter にはハッシュタグと呼ばれる機能がある．ハッシュタグには，ツイートにカテゴリをつけて検索しやすくするタグのような役割がある．しかし，テレビ番組名やイベント名などの共有したい話題をハッシュタグとする場合だけでなく,「おもしろハッシュタグ」と呼ばれるようなものも多数存在する．このようなハッシュタグ内に含まれる単語には，話題としての意味はないため，収集したすべてツイートからハッシュタグを除去する．

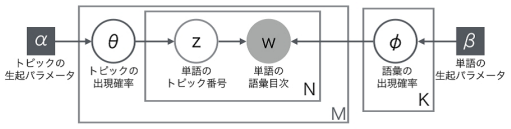


図 1　LDA のグラフィカルモデル

3.2 基底トピックの抽出

本研究では，3.1 節のデータクレンジングを行った後の，各ユーザのツイートを結合したものを一つの文書として特徴量算出を行う．特徴量の抽出のモデルとしては，bug-of-words（BoW）を利用する．

これまでに，文章ごとの特徴量算出の手法として様々な手法が提案されており，一般的に単語の出現頻度と希少性を考慮した TF-IDF を算出する方法が多く利用されている．TF-IDF では，文書中の一致する単語のみを考慮しているため，類似した意味の語の影響が無視される．本研究で分析対象とするテキストはツイートであるため，同じ内容を表す文章であっても口語表現や各ユーザの語彙による表記ゆれが発生する可能性が高いと考えられる．そこで，文章の背後に潜在的なトピックが存在すると仮定したソフトクラスタリングを行う手法であるトピックモデルを利用する．トピックモデルとしては，様々なものが提案されているが，本研究では，確率的トピックモデルである Latent Dirichlet Allocation（LDA）を利用する．

3.2.1 LDA

LDA は，Blei らにより提案されたトピックモデルである [8]．トピックモデルとは，ある文書 $\mathbf{w^d} = (w_1, w_2, ..., w_{N_d}) \in D$ が単一または複数のトピックに属する単語から構成されていることを仮定した言語モデルである．LDA ではこの仮定のもと，文書を構成するトピック分布 $\Theta^d = (\theta_1, \theta_1, ..., \theta_K)$ と，それらのトピックごとの単語生成分布 $\Phi^t = (\phi_1, \phi_2, ..., \phi_N)$ に基づき，確率的に文書を生成するモデルを構築する．ここで，変数 K はトピック数，M は文書数，N は各文書の単語数を表す．LDA の生成過程は次のようになっており，グラフィカルモデルで表したものを図 1 に示す．

1. トピック毎に，
   i. 単語分布を生成：$\Phi^t \sim Dir(\beta)$
2. 文書毎に，
   i. トピック分布を生成：$\Theta^d \sim Dir(\alpha)$
   ii. 単語毎に，
      a トピックを生成：$z^d_{dw} \sim Mult(\Theta_d)$
      b 単語を生成：$w_{dw} \sim Mult(\Phi_z)$

3.2.2 コーパス文書

トピックを抽出するためのコーパスとなる文書によって，作

表 1　除外対象とする品詞

Type	Subtype	Type	Subtype
名詞	動詞非自立的	形容詞	接尾
	数		非自立
	接続詞的		
	代名詞		
	非自立		
	特殊		

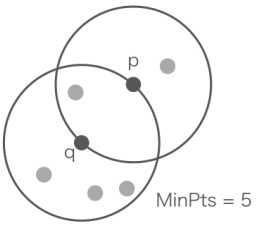


図 2　直接到達可能

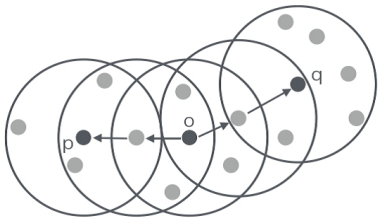


図 3　密度到達可能

成されるトピックには特徴が現れると考える．本研究では，以下の 3 種類の文書の BoW をコーパスとして用いる．

（1）AllAbout の記事

AllAbout は生活情報記事サイトであり，様々な分野の専門家が監修した記事を掲載している．AllAbout の記事は，住居，旅行，恋愛，マネー，グルメ，ビューティーのカテゴリに分類されており，多くのユーザの日常生活に関する話題を抽出できると考えられる．

（2）ランダムに抽出した Twitter ユーザのテキスト

ランダムに取得したユーザのツイートを文書とする．Twitter では情報の拡散が行われるために得られる流行的な話題が抽出できると考えられる．また，文書がユーザ単位であるため，一般的ではない趣味などの話題が抽出できると考えられる．

（3）Wikipedia

Wikipedia は一般ユーザが自由に執筆できる百科事典である．Wikipedia の記事からは，広く概念的な話題を抽出できると考えられる．

3.2.3　ストップワードの設定

3.2.2 節の文章のトピックを作成するために利用する BoW には，全ての品詞を使用するわけではない．ユーザの興味の対象物および対象物の属性を話題と考え，BoW を構成する単語は名詞と形容詞に限定する．名詞または形容詞であっても，代名詞，数および他の単語に結合して意味を添えるだけの単語は単体では意味をなさず，話題の特徴とはなり得ない．また，どのような話題であっても出現する単語もあらかじめ除去する必要がある．そこで，収集したすべてのテキストに対して次の処理を行う．

MeCab [11] を用いて形態素解析を行い，テキストを単語単位に分割する．本研究では，テキストに含まれる固有名詞や専門用語を扱えるようにするため，MeCab で利用する辞書に対して Wikipedia に存在する項目名の登録を行った．形態素解析後は，ユーザごとに単語の品詞が名詞または形容詞であると判定された単語のうち，表 1 に示した品詞を除外した単語のみを利用する．品詞の分類には，IPA 品詞体系を利用した．そして，全文書中における各単語を含む文書の出現頻度 Document Frequency（DF）を求め，この値が閾値以上のものは話題に関わらず出現する単語として除去する．

3.2.4　基底トピックによるユーザの特徴量算出

抽出した基底トピックを利用して，各ユーザのツイートの特徴量を算出する．具体的には，基底トピックを利用して，ユーザごとに各トピックに対する含有率を計算する．全ユーザに対して基底トピックの含有率を計算することで，各ユーザの基底トピック含有率の分布が得られる．この分布のベクトル表現をユーザの特徴量ベクトルとして定義する．

3.3　中心的なユーザの抽出

ユーザの特徴量ベクトルを利用し，ファン集合内のユーザのクラスタリングを行う．ファン集合内のユーザには，スパムアカウントなどファンと定義することが不適切な外れ値的なユーザも存在するため，全ユーザをクラスタリングするのではなく，中心的なユーザをグループ化することで，精度の向上が期待できる．そのため，密度に基づいたクラスタリング手法である Density-based spatial clustering of applications with noise（DBSCAN）[12] を用いる．

DBSCAN は密度準拠のクラスタリング手法である．密度準拠のクラスタリングは，近傍の最大半径 ε と近傍内に含む最小オブジェクト数 $MinPts$ の 2 つのパラメータを持つ．あるデータ点 p から距離 ε 以内にある近傍を $N_\epsilon(q) : \{p \in D | dist(p,q) \leqq \epsilon\}$ と定義し，以下の条件を満たすときに点 p から q へ直接到達可能であるといい，この状態を図 2 に示す．

（1）　$p \in N_\epsilon(q)$

（2）　$N_\epsilon(q) \geqq MinPts$

この定義のもと，DBSCAN では，図 3 の点 o から点 p，点 o から点 q のように直接到達可能な対象の集合で極大のものをクラスタとして抽出する．どのクラスタにも属さないオブジェクトはノイズとして扱う．

3.4　音楽アーティスト間の類似性

音楽アーティストの中心的なファン集合の特徴を利用して，音楽アーティスト間の類似度を定義することを考える．本手法では，ファン同士の類似度の全ての組み合わせの平均により音楽アーティスト間の類似度を定義する．形式的には，音楽アーティスト a_1, a_2 の類似性 $\mathrm{sim}(a_1, a_2)$ を以下の式 1 によって定義する．

$$\mathrm{sim}(a_1, a_2) = \frac{\sum_{u_1 \in F(a_1), u_2 \in F(a_2)} \mathrm{Cos}(\mathbf{t}(u_1), \mathbf{t}(u_2))}{\mid F(a_1) \mid \times \mid F(a_2) \mid} \quad (1)$$

ここで，$F(a)$ はアーティスト a の中心的なファン集合である．また，$\mathbf{t}(u)$ は，ユーザ u の基底トピックの分布を表すベクトルである．また，$\mathrm{Cos}(\mathbf{t}_1, \mathbf{t}_2)$ は，ベクトル $\mathbf{t}_1, \mathbf{t}_2$ のコサイン相関値を表す．

4.　予 備 実 験

提案手法における，LDA によるトピックモデル構築に利用するデータセットおよびパラメータ決定のために，2 つの予備

実験を行った．

4.1 実験環境

トピック抽出のためのコーパスとして利用する文書集合の取得時期と件数を示す．AllAbout の記事は，AllAbout 公式アカウント（@allabout_news）の 2017 年 1 月 26 日時点での全ツイート 29,088 件内で紹介された記事を重複なく抽出した結果の 19,295 件の記事である．ランダムなユーザのテキストは，UserStreamingAPI を利用して取得した．取得したユーザは，2017 年 2 月 9 日から 10 日までの 1 日間のツイート発信ユーザ 650,100 人から 100,000 人をランダムサンプリングし，重複を除外した結果の 94,483 人である．取得したユーザ毎の最新ツイート 200 件の集合をコーパスとして用いる．Wikipedia の記事は，2016 年 12 月 2 日時点の最新ファイルを利用した [14]．

4.2 LDA モデル構築のためのトピック数の検討

LDA では，文書から抽出するトピック数をパラメータとして与える必要があり，このパラメータにより抽出されるトピックおよびトピックモデルの性能が変化する．そのため，基底トピック抽出に利用する 3 種類のコーパスに対し，適切なトピック数を検討する．

4.2.1 実験方法

トピックモデルの評価指標として一般的には，Perplexity と Coherence [10] の 2 つが利用される．Perplexity はモデルの予測性能を評価するための指標である．Coherence は抽出されたトピックの品質（人間にとってわかりやすいか）を評価するための指標である．そのため，良いとされるトピックは Perplexity と Choherence のどちらで評価されたかで異なることもある．

本実験では，まず Perplexity の指標によりおおまかなトピック数の推定を行う．次に，推定されたトピック数で抽出したトピックの品質を評価して最終的にパラメータとして与えるトピック数を検討する．

Perplexity はモデルの予測（汎用）性能を測るための指標であり，式 2 で定義され，Perplexity が小さいほど性能が良いと言われている．

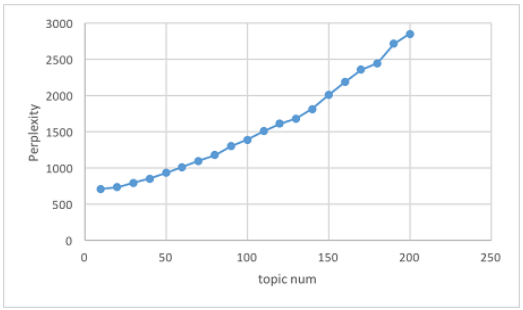
$$perplexity = \exp(-\frac{1}{N}\sum_{u}\log p(\mathbf{w}_u^{test} \mid M)) \qquad (2)$$

ここで，N はテストデータ中の全単語数，$\mathbf{w}_u$ はユーザ u のツイート集合に含まれる全単語である．M は確率モデル，$test$ はテストデータであることを表す．コーパス文書から 200,000 文書をランダムサンプリングし，8 対 2 の割合で学習用文書とテスト用文書に分割して Perplexity の計算を行い，トピック数 10 から 200 の間で Perplexity が低かった 5 件のトピック数を取得した．取得したトピック数をパラメータとして与えて抽出されたトピックを評価した．

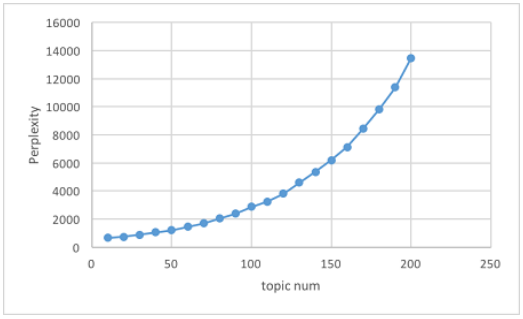
4.2.2 実験結果

トピック数を 10 から 200 まで 10 刻みで変化させて算出した Perplexity の値を図 4 に示す．横軸はトピック数，縦軸は Perplexity の値を示している．

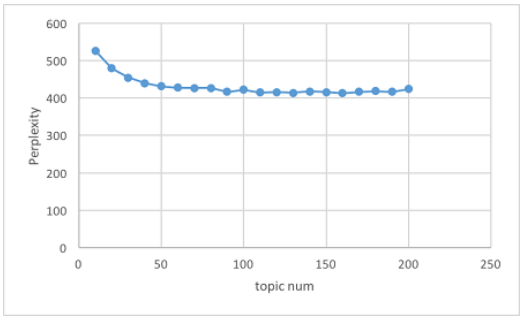
All Abou の記事およびランダムに取得したユーザのツイート文書においては，どちらもトピック数が大きくなるほど Perplexity が高くなる結果となった．トピック数を Perplexity が



[1] ランダムユーザのツイート



[2]All About 記事



[3]Wikipedia 記事

図 4　Perplexity 計算結果

低い値である 10, 20, 30, 40, 50 と変化させて抽出したトピックを評価したところ，トピック数が少ない場合には出現しない種類のトピックが存在することがわかった．特に，ランダムに取得したユーザのツイートは細かな分類でまとまったトピックが多数存在しているため，トピック数は多いほうが良いと判断した．

Wikipedia の記事においては，全体的に Perplexity は 3 桁であり予測性能は良いと言える．トピック数が 50 以下であるときは比較的高くなっているが，その後は変化がない．トピック数を 100, 120, 140, 160, 180, 200 と変化させて抽出したトピックを評価したところ，トピックの品質には差がないと判断した．

以上の結果より，本論文では LDA を利用する際に与えるトピック数は，All About の記事およびランダムユーザのツイートでは 50, Wikipedia の記事では 100 とする．

5. 基底トピック抽出のためのコーパスの検討

基底トピックを抽出するためのコーパスとして用いる文書の種類について検討する．生活情報記事サイト All About の記事，

ランダムに取得したユーザのツイート集合および Wikipedia の記事の 3 種類から LDA を利用してトピックを抽出し，基底トピックとして妥当性があるか評価を行う．

5.1 実験結果

All About は生活情報記事サイトであるため，ユーザの日常生活に関わる基本的な話題が抽出されることが期待できる．実際に抽出されたトピックを表 2，表 3 に示す．表中に示されている「トピック解釈」は，著者らがトピックの内容を解りやすく示すために付与したものであり，自動的に抽出されたものではない．トピックは，表 2 のように，All About の記事分類に沿った解釈ができる．さらに分類を細分化したような形のトピックも取得でき，表 3 には美容に関するトピックを抜粋している．スキンケア，ファッション，化粧などの美容に関係する様々なトピックにわかれている．また，全体的にトピックの解釈ができる単語集合が多く，上位単語にトピック解釈に沿わない単語が出現することは少ない．

ランダムユーザのツイートは文書がユーザ単位のツイートであるため，一般的ではない趣味や小規模な話題も抽出されることが期待できる．Twitter は即時性の高い情報源であるため，現実に起きている流行的な話題を抽出することが容易であると考えられる．実際に作成されたトピックを表 4 に示す．トピックの種類としては，ソーシャルゲームや恋愛などのユーザの趣味に関するグループ，広告や bot のグループ，感情を表すもの，時事ニュースなどが得られている．その他，解釈できないものや，複数の話題が混じっているようなトピックも存在している．

Wikipedia は広くジャンルを網羅していると言えるため，概念的な話題を抽出されることが期待できる．実際に作成されたトピックを表 5 に示す．トピックの種類としては，固有名詞などの細かな内容を表す単語ではなく誰もが知っている概念的な単語がトピックに含まれ，大きなまとまりでグループ化されている．ただし，キリスト教，仏教など個々の宗教ごとにトピックが形成されるなど，グループ化が細分化されているジャンルも少数ではあるが存在している．

検証の結果，コーパス文書から抽出できる基底トピックは，トピック内の単語の一般性に最も違いがあることがわかった．WIkipedia の記事，All About の記事，ランダムユーザのツイートの順に単語の一般度が高く，トピックのまとまりが広いものとなる．今回はユーザの特徴文書としてツイート集合を利用するため，大きなまとまりとしての単語は多く使用されないため，Wikipedia の記事から抽出したトピックでは個人の特徴の取得が難しい．ユーザの特徴文書と単語の一般性が近いコーパスは，ランダムに取得したユーザのツイートである．また，ユーザの特徴文書とするツイートの取得時期を合わせることで，同じ時期に，同じ話題に反応したという特徴を得ることができる．しかし，トピック内の単語に統一感がなく，解釈が難しいものが多いという問題点もある．人間が判断できるトピックの品質が良かったものは All About の記事である．よって，ユーザの特徴をより細かく抽出する際にはランダムに取得したユーザのツイート集合，誰もがわかる枠組みでユーザ層を抽出する際には All About の記事をコーパスに利用することが適切であると考えられる．

表 2　All About 記事トピック（記事分類）

トピック解釈	暮らし	マネー	健康	美容	恋愛	ビジネス	旅行
上位単語	掃除	お金	筋肉	アレンジ	恋愛	仕事	観光
	部屋	家計	体	ヘア	男性	言葉	温泉
	リフォーム	貯蓄	効果	髪	相手	相手	桜
	家	節約	エクササイズ	ヘアメイク	結婚	話	旅行
	洗濯	管理	姿勢	スタイル	男	関係	鎌倉
	トイレ	生活	運動	毛	コラム	ストレス	紅葉
	家事	財布	お腹	パーマ	関係	メール	名所

表 3　All About 記事トピック（美容）

トピック解釈	スキンケア	ファッション	化粧	産後ケア	ヘアアレンジ
上位単語	肌	ファッション	メイク	効果	ヘアスタイル
	美容	カラー	顔	ホルモン	スタイル
	ケア	小物	印象	産後	毛
	効果	デザイン	目	ビタミン	髪型
	乾燥	通販	色	ダイエット	パーマ
	対策	色	リップ	食欲	前髪
	スキンケア	バッグ	明るい	体	ミディアム

表 4　ツイートトピック

トピック解釈	政治	感情	ソーシャルゲーム	サッカー	グッズ
上位解釈	日本	笑	モン	選手	交換
	ニュース	ありがとう	スト	マジカルミライ	求
	韓国	やばい	白	サッカー	譲
	安倍	めっちゃ	猫	チーム	譲渡
	中国	かわいい	まとめ	大会	送料
	慰安	嬉しい	キャラ	試合	気軽
	政府	最高	運	横浜	検索

表 5　Wikipedia 記事トピック

トピック解釈	医療	メディア	アート	相撲	宗教	テクノロジー
上位単語	治療	放送	美術	力士	協会	開発
	疾患	番組	画家	部屋	宗教	技術
	感染	出演	画	場所	聖	システム
	頭痛	テレビ	浮世絵	大相撲	キリスト教	計画
	患者	ラジオ	絵	相撲	聖書	プロジェクト
	検査	制作	版画	敗	修道	設計
	症状	ツイッター	芸術	砂	聖女	コストパフォーマンス

6. 実　　験

音楽アーティスト間の類似度およびファン層の特徴を提案手法により抽出し，抽出結果についてアンケートにより印象調査を行い，提案手法の有効性および妥当性を評価した．

6.1 実験環境

分析対象として，表 6 の音楽アーティスト 20 名を用いる．これらの公式アカウントに対し，それぞれ 5,000 人のフォロワーをファンとし，各ユーザの 2016 年 12 月 4 日時点での最新ツイート 200 件を結合したテキストを文章集合として利用する．本実験で対象とした音楽アーティストはレコチョク (注4) のジャンル別ランキング上位から選択している．また，十分な量のデータが収集できないものは除外するため，公式アカウントのフォロワー数が 5,000 人以上のものに限定している．基底トピックは，4. 節で決定したコーパスとトピック数を与えて，All About 記事集合から作成したトピックを用いる．

クラスタリングに用いる DBSCAN のパラメータとして，近傍の最大半径 ε と近傍内に含む最小オブジェクト数 $MinPts$ を与える．ε には各音楽アーティストのフォロワー同士の類似

(注4)：http://recochoku.jp/

表 6　分析対象の音楽アーティスト

ジャンル	J-POP	邦楽ヒップホップ・R&B・レゲエ	邦楽ロック	ダンス・アイドル
アーティスト名	星野源	三浦大知	backnumber	Perfume
	西野カナ	AK-69	RADWIMPS	きゃりーぱみゅぱみゅ
	サザンオールスターズ	湘南乃風	ONE OK ROCK	Berryz 工房
	ナオト・インティライミ	SKY-HI	THE YELLOW MONKEY	℃-ute
	林部智史	DAOKO	L'Arc en Ciel	BABYMETAL

度の平均値の 1/3，$MinPts$ には各音楽アーティストのノイズ除去後のフォロワーの人数の 1/20 をパラメータとして与える．フォロワー同士の類似度は，ユーザごとに基底トピックの含有率を特徴ベクトルとしたコサイン類似度とする．2 人のユーザの特徴ベクトルをそれぞれ $\mathbf{a}$, $\mathbf{b}$ としたとき，コサイン類似度は次の式 3 で定義される．

$$\mathbf{a} = (a_1, a_2, \ldots, a_n)\ \mathbf{b} = (b_1, b_2, \ldots, b_n)$$

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\ \mathbf{a}\ ||\ \mathbf{b}\ |} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}} \qquad (3)$$

印象評価調査の対象は 20 代の男女 17 名であった．対象者には，ファン層の特徴とアイテム間の類似性に関するアンケートを実施した．ファン層の特徴としては，音楽アーティストのファン層における頻出トピックのトピック解釈を記述し，「ファンが興味のある（または好む）分野か」を○×で評価してもらった．アイテム間の類似性としては，音楽アーティスト間の類似度を「似ている」「やや似ている」「どちらとも言えない」「やや似ていない」「似ていない」の 5 段階で評価してもらった．

6.2　ファン層特有の基底トピックによる評価

6.2.1　実 験 方 法

評価のために，音楽アーティストのファン層特有の基底トピックを得る必要がある．多くのユーザにおいて横断的に出現しているトピックは特徴的なトピックではないため，ユーザのトピック分布に対してトピックの出現頻度による重み付けを行う．具体的には，各トピックにおいて式 4 に示す ITF(Inverse Topic Frequency) を算出し，ユーザのトピック含有率に掛け合わせる．

$$ITF(t) = \log \frac{N}{TF(t)} + 1 \qquad (4)$$

ここで，$TF(t)$ は，あるトピック t が出現する文書の数であり，N は総文書数を表す．クラスタリングにより抽出されたグループごとに，重み付けを行ったトピック分布の平均を求め，上位トピックから妥当性について検討する．

妥当性の評価として，音楽アーティストのファン層に好まれる分野に関するアンケートを実施して比較を行う．アンケートには，ファン層における頻出上位トピック 3 件のトピック解釈を記載し，「音楽アーティストのファンが興味がある（または好む）分野か」の評価を得る．評価結果は 1 または 0 の点数として扱う．

6.2.2　実 験 結 果

アンケートによる評価点が 0.5 以上を正解データとした場合，20 名の音楽アーティストごとに頻出トピックの上位 3 件を取得した合計 60 件中の正解数は 36 件となり，適合率は 0.6 である．

表 7　頻出トピック：℃-ute

トピック解釈	恋愛・結婚	アイドル	ランチ
上位単語	恋愛	名前	カフェ
	男性	アイドル	店
	相手	丼	東京
	結婚	番組	メニュー
評価値	0.7778	1.0000	0.6667

表 8　頻出トピック：サザンオールスターズ

トピック解釈	オリンピック	イベント	海外旅行
上位単語	選手	東京	旅行
	企業	ホテル	夜景
	発表	開催	海外
	五輪	イベント	利用
評価値	0.7273	0.7273	0.7273

上位 3 件のトピック全ての評価点が 0.5 以上となった音 g 買うアーティストは 5 名である．特に評価点が高い「サザンオールスターズ」および「℃-ute」の頻出基底トピックを表 7，表 8 に示す．

「℃-ute」のファン層では「恋愛・結婚」「アイドル」「ランチ」に関するトピックに含まれる単語の含有率が高い．「℃-ute」が女性アイドルグループであることからファンのツイートに「アイドル」に関するトピックが頻出することには妥当である．「恋愛・結婚」「ランチ」に関しては，ファンが年配の男性だけでなく若い女性も多いことから評価点を高くしたという意見が多かった．「サザンオールスターズ」のファン層では「オリンピック」「イベント」「海外旅行」に関するトピックに含まれる単語が頻出している．全体的に年齢層が高い，安定した収入があるなどの特徴を持つ人々に好まれるというイメージが一致したため評価点を高くしたという意見が多かった．

ファン層特有の基底トピックの妥当性は，全体的な適合率から見ると良い結果ではない．しかし，数名の音楽アーティストに関しては，ファンの年齢層や性別による特徴が現れたトピックが上位となり，評価点も高いという結果が得られた．このことから，ファンの特性として日常的な話題を用いることは有効であると考えられる．

6.3　アーティスト間類似度による評価

6.3.1　実 験 方 法

対象の公式アカウントをフォローしているユーザごとにグループ化し，グループ同士のユーザ間類似度の平均値を音楽アーティスト間類似度と定義する．音楽アーティスト a_1，a_2 の間の類似度は式 1 を利用した．

ピアソンの相関係数を利用して，評価値と計算値の相関を評

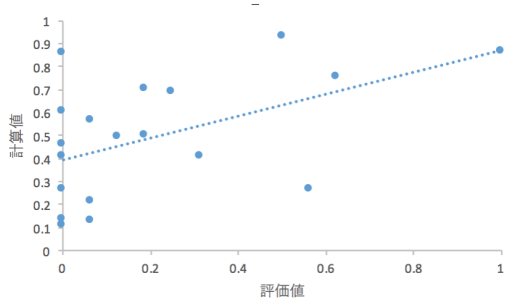


図 5　類似度の評価値と計算値の相関

表 9　「℃-ute」と他の音楽アーティスト間の類似度

	アーティスト名	ジャンル	類似度	評価値
上位	きゃりーぱみゅぱみゅ	ダンス・アイドル	0.5034	0.5500
	Berryz 工房	ダンス・アイドル	0.5030	0.9500
	星野源	J-POP	0.5000	0.0000
下位	BABY METAL	ダンス・アイドル	0.4056	0.6667
	L'Arc en Ciel	邦楽ロック	0.3879	0.0200
	THE YELLOW MONEKY	邦楽ロック	0.3514	0.0000

価する．評価値には，音楽アーティストのファン層に好まれる分野に関するアンケートの結果を利用する．アンケートでは，音楽アーティストの組み合わせの類似性について５段階評価を得る。評価値は０～１に正規化する．

6.3.2　実 験 結 果

分析対象とした音楽アーティストのひとつである「℃-ute」について，他 19 の音楽アーティストとの類似度の散布図を図 5 に示す．横軸を評価値，縦軸を計算値としている．近似直線は右上がりであり，正の相関があると言える．相関係数 $|r| = 0.4447$, p 値 $= 0.0667$ である．提案手法により算出された類似度は，アンケートによる印象評価の結果との間にやや強い正の相関を持つことが明らかになった．

類似度の計算結果の上位下位それぞれ３件を表 9 に示す．ジャンルが同じ「ダンス・アイドル」である音楽アーティストが上位に現れている．しかし，全体的に類似度の差が小さいため，上位下位にあまり意味がないとも考えられる．類似度の上位アーティストのトピック分布は，比較的「ファッション」に関するトピックの含有率が高くなっているが，同じファッション分野でも「きゃりーぱみゅぱみゅ」ならば原宿系,「星野源」ならばサブカル系と好みは異なると考えられる．

以上の結果から，All About の記事から基底トピックを抽出した場合，音楽アーティストのジャンルなど，大まかなまとまりは推定することができるが，詳細なファン層の推定は難しいことがわかった．

7. おわりに

本論文では，Twitter 上における日常的な話題から音楽アーティストのファン層と抽出する手法を提案し，話題を表すトピックの元となるコーパスの検証を行い，各コーパスから得られるトピックに基づいたユーザの特徴量からファン層を抽出した．得られたファン層に関して，アンケートによる印象評価結果をもとに，音楽アーティストごとの特有の話題および音楽アーティスト間の類似度の２点から提案手法の有用性の検証を行った．検証の結果，音楽アーティストごとに基底トピックの分布には印象評価と一致する特徴が現れ，日常的な話題がファン特性の抽出に有効であることがわかった．しかし，音楽アーティスト間類似度には大きな差が出ず，類似度上位の組み合わせとしてもジャンルや年齢層以外の結びつきは確認できなかった．理由としては，基底トピックの分布から広いまとまりでの興味・関心のある分野は抽出できているが，その分野におけるユーザの好みなどが考慮できていないことが考えられる．

今後は，複数のコーパスから抽出した基底トピックを組み合わせ，一般性の高い話題から詳細な話題まで絞り込みを行うことで精度の向上を目指す．また，ツイート文書における話題以外の要素として，語尾や絵文字の特徴も加味し，SNS におけるファンの特性の抽出方法を検討していく予定である．

文　　献

[1] 総務省：平成 28 年版情報通信白書, 第 2 部, 第 4 章, 2016.
[2] 池田和史, 服部元, 松本一則, 小野智弘, 東野輝夫, ”マーケット分析のための Twitter 投稿者プロフィール推定手法”, 情報処理学会論文誌コンシューマ・デバイス&システム,vol.2,No.1,pp.82-93,2012.
[3] 西村章宏, 土方嘉徳, 三輪祥太郎, 西田正吾, ”一般ユーザの観点に基づく Twitter からの人物関係の可視化と事例の考察”, 情報処理学会論文誌,vol.56,no.3,pp.972-982,22.
[4] 田原如菜，坂地泰紀，酒井浩之，”Twitter からのキャラクター印象表現の抽出”, 信学技法,vol.56,no.3,pp.972-982,22.
[5] 長浜祐貴, 遠藤聡志, 當間愛晃, 赤嶺有平, 山田考治, ”Twitter の投稿文書による人物像の推定”, 情報処理学会第 76 回全国大会,2014.
[6] Twitter Developers, ”Twitter Developer Documentation”, https://dev.twitter.com/docs (2017-01-23)
[7] Koga H, and Taniguchi T. ”Developing a user recommendation engine on twitter using estimated latent topics”, Proc. HCII (The 14th International Coference on Human-computer Interaction), pp.461-470, 2011.
[8] David M Blei, Andrew Y Ng, and Michael I Jordan. ”Latent Dirichlet Allocation”, the Journal of machine Learning research, Vol. 3, pp.993-1022, 2003.
[9] 佐藤一誠. 2015. トピックモデルによる統計的潜在 意味解析. 東京：コロナ社, 2015. 4339027588.
[10] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. ”Reading Tea Leaves:How Humans Interpret Topic Models”, Neural Information Processing Systems, pp9, 2009.
[11] MeCab - Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/
[12] M.Ester, H.-P.Kriegel, J.Sander, and X.Xu ”A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, KDD1996
[13] Ankerst, Mihael., Breuning, Markus M., Kriegel, Hans-Peter., Sander., Jorg. (1999) ”OPTICS: Ordering Points To Identify the Clustering Structure,”ACM SIGMOD Int. Conf. Management of Data, Philadelphia, PA.
[14] Wikimedia Downloads, https://dumps.wikimedia.org/jawiki/