# Twitterを用いた音楽の鑑賞行動とトレンドの分析
# Analysis of Music Listeners' Behaviors and Trends by Using Twitter

山口　由馬
Yuma Yamaguchi
法政大学情報科学部ディジタルメディア学科
*E-mail: yuma.yamaguchi.2h@stu.hosei.ac.jp*

## Abstract

*Sales of music CDs are becoming less and less standard as the popularity of music due to the presence of purchase bonuses and the decrease of people who purchase CDs. In this paper, we propose utilizing Twitter as a new source of information to get to know the behaviors of music listeners and the trends of music. For our analysis, we collect tweets by using the prevalent hashtag #NowPlaying. Specifically, automatically extracting the names of artists and songs from collected tweets and also combining them with additional information including times, positions, and user terminals associated with tweets, we analyze the classification of music genres, the time periods of music listening, and the influence of music-related events. Using social media to process behaviors of music listeners and trends of music, we attempt to analyze naturally arising trends that are hard to get to know from sales of CDs. As a result, it was able to visualize when people were listening to music, but it turned out that analysis using positional information was difficult. Also, it turned out that there were old musical pieces that people continuously listened to as well as musical pieces that people listened to under the influence of the media, which were different from the actual sales of CDs.*

## 1. はじめに

近年，定額制の音楽配信サービスの登場による CD 購入者の減少や CD の特典商法の存在により CD 販売枚数が音楽の流行を表す基準ではなくなってきている．本研究では音楽のトレンドを知るための新たな情報源として Twitter を利用することを提案する．特に Twitter 内で流行しているハッシュタグ#NowPlaying（リアルタイムで聴いている曲を投稿する時につけるタグ）を利用してツイートを収集し，音楽の鑑賞行動とトレンドを分析する．より具体的には，ツイートに書かれているアーティスト名と曲名を自動抽出して，時刻，位置，発信端末といった追加の情報と組み合わせることで，音楽ジャンルの分類や，音楽鑑賞の時間帯を分析する．本研究ではこのようにソーシャルメディアを利用して，音楽の鑑賞行動やトレンドを統計的に処理することによって，CD の購入実績では判断出来ない自然的に発生したトレンドの分析を試みる．その結果，音楽が聴かれている時間帯を抽出することはできたが，位置情報を利用して分析することは現状では難しいことがわかった．また，メディアの影響でトレンドになる楽曲と新曲でなくても継続的に聴かれている楽曲があり，CD の購入実績と異なった結果が得られた．

## 2. 分析手法

### 2.1. 概要

本研究では Twitter を使用し，ツイートされている内容からアーティスト名と曲名を抽出することにより，音楽の鑑賞行動やトレンドを分析する．これまでにもテレビ番組を対象として Twitter の情報を利用する研究 [1] [2]やハッシュタグの研究 [3] [4]は存在している．本研究の特徴は，ツイートから得ることのできる位置情報や時刻など様々な情報を収集しそれを複合的に使うことである．

### 2.2. ツイートの収集

#NowPlaying をつけてリアルタイムでアーティスト名と聴いている曲名をツイートする Twitter 内の流行に着目して，ツイートの収集を行う．また，Twitter からしか得ることのできない情報を総合的に分析していくために，ツイートだけでなく，言語，位置情報，リツイート，投稿時間，発信端末を同時に取得しておく．

### 2.3. 鑑賞行動の分析

本研究では#NowPlaying を付けてツイートしている人の鑑賞行動を分析する．具体的には，位置情報をつけツイートしている人とツイートされる時間帯に着目し，音楽を聴いている時の行動を分析する．

### 2.4. トレンドの分析

ツイートからは得られない楽曲のジャンルなどの知識を分析に用いるために，事前に楽曲データベースを作成する．取得したツイートに対しては，テキスト処理を行うことで楽曲を抽出する準備を行った後，楽曲データベースとテキスト処理されたツイートを照合することで楽曲を抽出する．この抽出された楽曲のジャンルを分類した上で，頻出する楽曲やアーティストのデータを分析する．

#### 2.4.1. データベース作成

アーティスト名と曲名を照合するために楽曲データベースを用意しておく．本研究では，著者の iTunes に入っているアーティスト名をキーワードとして iTunes Store API [5]を利用し，ジャンル，発売年，曲名を取得し，ア

Supervisor: Prof. Hiroshi Hosobe

ーティスト名とともに楽曲データベースとして保存しておく．この楽曲データベースのアーティスト名と曲名を形態素に分け，tf-idf [6] を計算する．形態素から楽曲を紐付けしておき，ツイートから楽曲を照合するための形態素データベースを作成しておく．

#### 2.4.2. テキスト処理

2.4.1 節に述べた方法で収集したツイートからアーティスト名と曲名を自動抽出するために，ツイートから不要なテキストを削除していく．最初に正規表現で Twitter の特徴的な語句を削除する．正規表現を用いた特徴的な語句の削除では，引用リツイート，ハッシュタグ，メンション，URL や，特定のアプリケーションから Twitter に投稿された時に付随する，“by”や“なうぷれ”といった単語も削除する．

#### 2.4.3. 楽曲照合

楽曲データベースに保存されている楽曲をツイートから照合する．最初に前項で得られたテキストに対して形態素解析を適用する．アーティスト名と曲名の形態素データベースとツイートの形態素を照合し，形態素の前後関係も利用しながら楽曲の範囲を絞り，絞られた楽曲の形態素とツイートの形態素をさらに照合することで，ツイートに楽曲データベースの楽曲が含まれているか調べる．一般には，常に楽曲が 1 つに絞られるとは限らず，複数の楽曲が残ってしまう場合がある．本アルゴリズムではこのような場合に，楽曲の抽出に失敗したと見なし，楽曲を出力しないこととしている．

#### 2.4.4. 分析

照合した楽曲のジャンルを分類する，最も多い割合のジャンルから多く抽出されたアーティスト上位 5 名の出現回数，ユーザ数を算出しておく．ツイートの中には同ユーザが複数回ツイートをして出現回数を増加させていることが考えられるので，そのような重複を除き数えたものがユーザ数である．これらの統計的なデータを使用することで，販売枚数では分からない音楽のトレンドを分析する．

## 3. 実装

### 3.1. ツイートの収集

本研究ではプログラムの実装に Python を使用した．Twitter からツイートを収集するのに，Twitter API のライブラリ Tweepy [7]を使用した．10 月 2 日から 12 月 2 日の期間に渡り，リアルタイムのデータを Streaming 機能で収集した．ツイートの内容だけでなく位置情報，時間，リツイート数のデータも保存した．

### 3.2. テキスト処理

本研究では Python の正規表現モジュール [8]を用いてツイートから特徴的な語句を削除していく．

### 3.3. 形態素解析

前項で正規化されたツイートを MeCab [9]を使用して形態素に分ける．形態素に分ける際，照合の実行速度と照合率を上げるために，名詞，形容詞，動詞のみを保存する．

## 4. 予備実験

### 4.1. 方法

本実験で使用する閾値を決定するために，ランダムに 100 件のツイートを選択し，その 100 件を対象に照合率と実行時間を求める予備実験を行った．本研究の楽曲照合アルゴリズムは楽曲を誤って出力する場合がある．また，楽曲データベースの作成にあたり，著者自身の iTunes のデータに含まれるアーティストのみを対象としているため，ツイートが参照している楽曲が楽曲データベースに入っていない場合がある．これらを考慮し，本予備実験では楽曲照合の結果を分類(表 1)して調べた．

表 1 楽曲照合の分類

<table>
<tr><td rowspan="2">A: 楽曲データベースの中にある楽曲を出力した場合</td><td colspan="2">Aa: 出力された楽曲が正しかった場合</td></tr>
<tr><td>Ab: 出力された楽曲が誤っていた場合</td><td>Ab1: ツイートが参照している楽曲がデータベースの中にあるのに，別の楽曲を出力した場合.<br>Ab2: ツイートが参照している楽曲がデータベースの中にないのに，楽曲を出力した場合.</td></tr>
<tr><td rowspan="2">B: 楽曲データベースの中にある楽曲を出力しなかった場合</td><td colspan="2">B1: ツイートが参照している楽曲がデータベースの中にあるのに，楽曲を出力しなかった場合</td></tr>
<tr><td colspan="2">B2: ツイートが参照している楽曲がデータベースの中になく，楽曲を出力しなかった場合.</td></tr>
</table>

形態素は全体で 381,142 個あり，そのうち tf-idf が 1.0 以上のものは 87,527 個であった．一方，tf-idf の平均値は 0.41 であり，平均値以上の形態素は 193,976 個であった．

### 4.2. 結果

表 2 に照合率を，表 3 に実行時間を示す．

表 2 照合率

| 照合率 | Aa [%] | Ab1 [%] | Ab2 [%] | B1 [%] | B2 [%] | Aa/(Aa+Ab1 +Ab2) [%] |
|---|---|---|---|---|---|---|
| 閾値 1.0 | 22 | 5 | 10 | 23 | 40 | 59.5 |
| 閾値を全体の平均値とした場合 | 29 | 8 | 15 | 13 | 32 | 55.8 |

表 3 実行時間

| 実行時間 | 100 ツイートあたりの時間[秒] |
|---|---|
| 閾値 1.0 | 331.4 |
| 閾値を全体の平均値とした場合 | 351.3 |

### 4.3. 考察

**照合率．**予備実験の結果，閾値を低く設定すると，出力されない割合(B1, B2)が低くなり，正しい出力(Aa)の割合は高くなるが，誤った出力(Ab1, Ab2)の割合も高くなった．一方で，楽曲が出力された場合にそれが正しい割合(Aa/(Aa+Ab1+Ab2))は，閾値を高く設定すると高くなった．本実験では出力なし(B1, B2)のものは使われないが，B1 の割合が低ければ，楽曲データベースに含まれる楽曲の取りこぼしが少ないと言え，理想的である．

**実行時間.** 楽曲を照合するための形態素の tf-idf の閾値によって実行時間が変化した．閾値が低くなるほど形態素の数が多くなるため，実行時間が長くなった．逆に閾値を高くするほど実行時間は短くなった．

**結論.** 予備実験の結果を踏まえて，本実験では閾値として tf-idf の平均値を用いる．楽曲が出力された場合にそれが正しい割合(Aa/(Aa+Ab1+Ab2))が高くなることから，閾値として 1.0 を用いることも考えられる．しかし，取りこぼし(B1)が少なく，多くの楽曲を抽出できていることから，tf-idf の平均値を採用するものとする．

## 5. 実験

鑑賞行動の分析とトレンドの分析の 2 種類の分析を行った．鑑賞行動の分析は音楽を聴く時間帯と位置情報について行う．トレンドの分析では音楽のジャンルで分類し頻出度の高いアーティストをランキング化する．

### 5.1. 鑑賞行動の分析

日ごとのツイート数を図 1 に示す．休日の鑑賞行動に特殊性が見られ，ツイートが増加する場合と減少する場合の両方が存在した．また，天気との関連を調べたところ，晴れや雨に関係なくツイート数が増減していた．

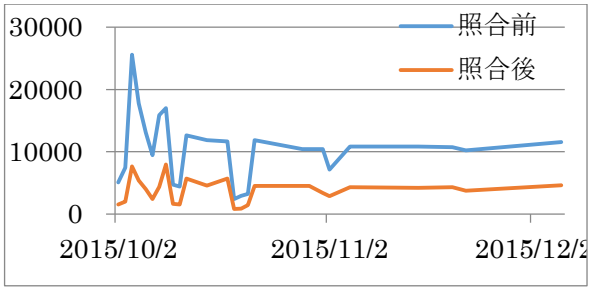


図 1 日ごとのツイート数

曲を聴く時間帯を分析するために，最もツイート数の多かった 10 月 4 日のツイートに対して楽曲の自動照合を行い，時間ごとのツイート数を図 2 に示す．朝の通勤時間帯や夕方の帰宅時間帯にツイートする人が多いと予想していたが，図 2 によると 13:00～15:00 の昼休みや 22:00～24:00 の就寝前の時間帯に曲を聴きながらツイートする人が多いことが分かった．

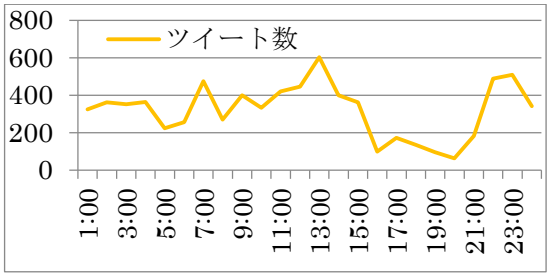


図 2 時間ごとのツイート数

本研究で収集したツイート 231,240 件のうち位置情報が付けられたツイートは 408 件であり，全体の 0.176%であった．位置情報を地図上に表し分析した結果，ハッシュタグ#NowPlaying の使用者は関東圏に集中していた．

この 408 件のデータを自動抽出しジャンル分けした結果を図 3 に示す．分析の結果，アニメ・ゲームが大半を占めていた，その理由として Forsquare という位置情報のアプリケーションを使用し，一駅ずつアニメの音楽を聴きながら位置情報を付けツイートする人がいることが分かった．

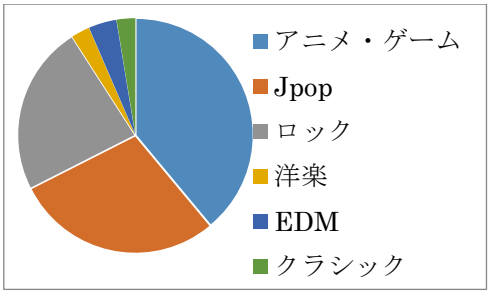


図 3 位置情報ツイートのジャンル分け

本研究では，位置情報から地域別での音楽ジャンルの違いなどを分析することを目標としていたが，#NowPlaying と位置情報の両方を付ける人が少ないことや，特殊なアプリケーションを使用している人いることにより，ジャンルの偏りや正確な位置情報が取得できないことから，位置情報を利用して分析することは現状では難しい．

### 5.2. トレンドの分析

自動抽出した音楽データをジャンル分けした結果を図 4 に示す．ジャンルの 45%を占めていた J-Pop の中から多く抽出されたアーティスト上位 5 名を表 4 に示す．多く抽出された原因を分析することで，Twitter で観察される音楽のトレンドを理解する．

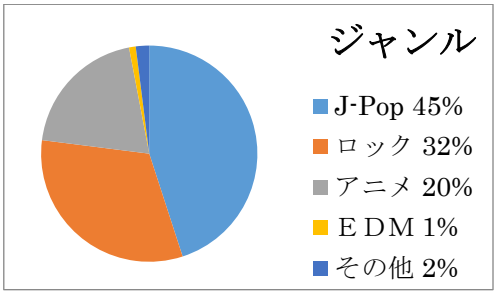


図 4 抽出した曲のジャンル分け

表 4　頻出 J-Pop アーティスト

| 順位 | アーティスト | 出現回数 | ユーザ数 |
|---|---|---|---|
| 1 位 | back number | 5505 | 3623 |
| 2 位 | Perfume | 4491 | 2945 |
| 3 位 | 福山雅治 | 2973 | 2032 |
| 4 位 | AKB48 | 2709 | 1201 |
| 5 位 | 宇多田ヒカル | 2624 | 1843 |

1 位の back number の「クリスマスソング」が特に多かった．原因としてこの楽曲が「月 9」の主題歌に採用されていることが考えられる．

2 位の Perfume に関しては特定の楽曲ではなくさまざまなものが聴かれていたが，10 月 30 日以降ツイート数が多くなった．理由として，前日にテレビで特集されたことが考えられる．

3 位の福山雅治は結婚の報道があってからツイート数が伸びた．報道以前は新曲「I am a Hero」のツイートが多かったが，報道後は「桜坂」など以前の楽曲が幅広く聴かれていた．テレビ番組や結婚報道などの影響により聴かれる曲が移り変わることが分かった．

4 位の AKB48 は新曲のツイート数も多かったが，最も多かったのは 2010 年に発売された「ヘビーローテーション」であった．

5 位の宇多田ヒカルに関しては，最近はメディアの露出がないにもかかわらず，「First Love」がよく聴かれていた．

トレンドの分析の結果，メディアに取り上げられてツイート数が伸びて瞬間的にトレンドとなるものと，新曲でなくても継続的に聴かれているものの 2 種類に分けられることが分かった．このようなトレンドは CD の販売枚数では得ることのできないデータであると言える．

## 6. 議論

本研究では，楽曲データベースを用いた楽曲照合アルゴリズムを提案し，その性能に関する予備実験を行った．しかし，これらは以下の理由により十分であるとは言えない．

- 誤った出力が多く存在する．Ab1, Ab2 が出現した理由として考えられるのは，形態素のつながりを完全には確認していなかったためである．この時点で楽曲の候補は絞られているため，多くの場合，形態素のつながりを十分に確認すれば，楽曲を 1 つに特定できたはずである．また，これによって B2 をより大きくすることもできたはずである．
- 予備実験で比較する閾値が少ない．予備実験では，2 つの閾値を比較してどちらを使うか決めた．一般には，閾値を上げると楽曲が出力されない割合が大きくなるが，実行時間は短くなると考えられ，これらの間にはトレードオフの関係がある．適切な閾値を決めるためには，比較する閾値を増やすことが必要である．
- 予備実験に用いるツイートが少ない．予備実験では，ランダムに選んだ 100 件のツイートに対して照合率を示したが，100 件で十分であったかどうかの検討は行っていない．しかし，一般にはツイート件数が多いほど適切な評価ができると言える．

これらの楽曲照合アルゴリズムと予備実験の問題を改善することで，より適切な分析が可能であると考えられる．

## 7. 終わりに

本研究では，Twitter からハッシュタグ#NowPlaying を使用してツイートを収集し，アーティスト名と曲名を自動抽出して，音楽の鑑賞行動とトレンドを分析した．今後，この分析手法をさらに活用していくために，邦楽だけでなく洋楽や，英語圏の国のツイートに対応することが必要だと考えられる．そのために以下の 3 つの改善すべき点が挙げられる．

- 本研究のテキスト処理は日本語を前提としていたため英語に対応できず，邦楽が主な照合対象になってしまったこと．
- ツイートの誤字脱字に対応する曖昧照合を実装しなかったこと．#NowPlaying をつけてツイートする人のほとんどが自動ではなく手動でアーティスト名と曲名を入力しているため，特に英語の綴り誤りが見られた．
- 楽曲データベースの規模が十分でなかったこと．単に楽曲数が不足しているのではなく，カバーされていないジャンルもあった．

これら 3 つを改善することにより，Twitter を利用した音楽トレンドの分析がさらに活用されるものになると考えられる．

## 文　献


[1] 山本裕輔, "テレビ番組に対する意見をもつ Twitter ユーザのリアルタイム検出(修士論文)," 早稲田大学基幹理工学研究科, 2013.

[2] 古宇田悠輔, "マイクロブログを用いたテレビ番組のリアルタイムなランキング手法(卒業論文)," 法政大学情報科学部, 2015.

[3] Z. Ma, A. Sun and G. Cong, "On Predicting the Popularity of Newly Emerging," *Journal of the Association for Information Science & Technology*, vol. 64, no. 7, pp. 1399-1410, 2013.

[4] S. Carter, M. Tsagkias and W. Weerkamp, "Twitter Hashtags: Joint Translation and Clustering (poster)," *Proc. ACM WebSci'11*, pp. 529:1-3, 2011.

[5] "iTunes Store API," [Online]. Available: http://www.apple.com/itunes/affiliates/resources/documentation/itunes-store-web-service-search-api.html.

[6] 秋葉拓哉, "TF-IDF で文書内の単語の重み付け," [Online]. Available: http://takuti.me/note/tf-idf/.

[7] "Tweepy: An easy-to-use Python library for accessing the Twitter API," [Online]. Available: http://www.tweepy.org/.

[8] "正規表現モジュール," [Online]. Available: http://docs.python.jp/2/howto/regex.html.

[9] "MeCab: Yet Another Part-of-Speech and Morphological Analyzer," [Online]. Available: http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html?sess=3f6a4f9896295ef2480fa2482de521f6.