|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| テーマ番号 | 1EP068 | |  | | |
| プロジェクト テーマ | 和文 | Twitterを用いた各都道府県のトレンドの解析 | | 指導教員 | 元木 光雄 准教授 |
| 英文 | Analysis of trends in each prefecture using Twitter | |
| プロジェクト メンバー | | 4EP2-30 髙橋 史弥 (Fumiya Takahashi) | | | |

People use Twitter to express their opinions and thoughts on matters and share news, event information, and so on. By collecting such information, we can know the trend at any time. In fact, Twitter show a list of trends. However, such trends are nationwide trends, thus it is not easy to know the trends of the specific region. The nationwide while trends are important, we consider that region trends are also important because they lead to the economic promotion and tourism promotion of the region as well. Therefore, in this study, we extract trends using tweets for each prefecture. However, if one just extracts trends from tweets by prefecture, there is a possibility that a nationwide trend will also be extracted. Therefore, we compare the trend by prefecture and the nationwide trend, and extract the trends of the region unique.

co-occurrence network，NPMI，maximal clique enumeration，data polishing

# まえがき

現在，スマートフォンやタブレット端末などの電子機器の急速な発展と普及により，若年層から高年層まで幅広い人々がSNS（ソーシャルネットワーキングサービス）を利用している．

日本では特にTwitterの利用者数が多い．実際に2019年4月時点で，日本国内の月間アクティブアカウント数は4500万人となっている[1]．Twitterでは公的機関や企業，有名人から一般人まで様々な人がツイートしている．したがって，Twitter利用者はテレビやラジオなどで報道される前の話題や報道されないような小さな話題まで知ることが出来る．この発信される話題は人々の暮らしや生活に大きく関わっている．しかし，Twitterは日本のトレンドしか公開していないため，身近な話題を知ることが出来ない．そこで，地域ごとでツイートを収集し解析することで，より自分自身に関係がある話題が入手できると考えた．

本研究では，Twitterから得られる位置情報付きツイートを用いて，都道府県ごとでトレンドの解析を行う手法を提案する．

# 提案手法

提案手法の流れを以下に述べる．まず，全国の位置情報付きツイートを収集する．次に，取得したツイートを都道府県ごとに分類する．さらに，ツイートから頻出単語集合を抽出し，それをトレンドとする．しかし，このままでは都道府県ごとの特徴的なトレンドとは言えない．そこで、都道府県のツイートから抽出した頻出単語集合と，その都道府県以外のツイートから抽出した頻出単語集合を比較し，都道府県ごとの特徴的なトレンドを抽出する．

## ツイートの取得と都道府県による分類

まず，TwitterのStreaming APIを利用して全国の位置情報付きツイートを収集する．Streaming APIの特徴は，リアルタイムのツイートが取得できることである．

次に，収集した位置情報付きツイートを都道府県ごとに分類する．そのために，Yahoo!リバースジオコーダAPIとYahoo!ジオコーダAPIを用いる．収集した位置情報付きツイートには，そのツイートがどこで発信されたか，緯度・経度または，バウンディングボックスという形で保存されている．緯度・経度で指定されている場合，それをYahoo!リバースジオコーダAPIに渡し，都道府県を一意に特定する．バウンディングボックスで指定されている場合，Yahoo!ジオコーダAPIに渡し，そのバウンディングボックス内の市町村を収集する．そして，その市町村が含まれる都道府県をツイートの発信元とする．本研究では，市町村データから複数の都道府県が導き出された場合，そのツイートは導き出された都道府県全てのデータとする．しかし，5つ以上の都道府県が導き出された場合，つぶやかれた範囲が広すぎると判断しそのツイートは除外する．

## 共起ネットワークの生成

まず，2.1で分類した都道府県ごとのツイートに対し，形態素解析エンジンMeCab [2]を用いて形態素解析を行い,単語単体で意味のある名詞，動詞，形容詞，形容動詞を抽出する．抽出した単語を頂点とし，同じツイート内に含まれる単語同士を辺で結んだ共起ネットワークを生成する．その際，各頂点にはその単語が含まれるツイート数を付与する．さらに，各辺には両端点の単語が同じツイートに含まれる回数を付与する．

次に，生成した共起ネットワークに対し，単語間の正規化自己相互情報量(NPMI: Normalized Pointwise Mutual Information)を計算する．NPMIの計算式を以下に示す．

ここで，は単語の出現頻度を表す．具体的には単語を含むツイート数の全ツイート数に対する割合である．は単語が同時に出現する頻度を表している．NPMIは，1.0から1.0の値をとり，0から1.0になるにつれて，2つの単語が同時に出現しやすい傾向にある．単語間の関係性の低い辺を削除するために，その閾値を指定し，小さいNPMIの値を持つ辺を削除することで共起ネットワークの二値化を行う．

## 都道府県ごとの特徴的なトレンド抽出

共起ネットワークの極大クリークをトレンドとする．

まず，2.2で生成した共起ネットワークに対して，極大クリーク列挙を行うと，類似した極大クリークが膨大になる．そのため，データ研磨という手法を用いて，類似した極大クリークを減少させる必要がある．データ研磨とは，共起ネットワークのすべての頂点ペアについて，それぞれの頂点の近傍の類似度が，あらかじめ指定した閾値より大きければ辺を接続し，小さければ辺を切断することで，新たな共起ネットークを生成する手法である．この処理を変化がなくなるまで繰り返す．これにより，共起ネットワークに存在している類似したクリークが併合される．

最後に，データ研磨を行った共起ネットークに対して，極大クリーク列挙を行うことで，都道府県ごとのトレンド（単語集合）を抽出することができる．しかし，この段階では，全国的なトレンドの可能性もあるため，都道府県ごとの特徴的なトレンドとは言えない．そのため，その都道府県以外のツイートデータに対しても，同じ処理を行い，単語集合を抽出する．そして，その単語集合と都道府県の単語集合を比較し，都道府県側の単語集合で同じ単語が含まれる確率が80％を超えた場合，同じトレンドとみなし，その単語集合は削除する．こうすることで，都道府県ごとの特徴的なトレンドを抽出する．

# 実験による評価

2019年11月30日20時から22時の間につぶやかれた位置情報付きツイートを実験対象とした．収集したツイート数は32,939件である．これを都道府県ごとに分類すると，1つのツイートが複数の都道府県のツイートに含まれることがあるため，32,972‬件と合計ツイート数が増大した．表1に，都道府県ごとのツイート数を示す．本稿では，最も取得ツイート数が多かった「東京都」のトレンドの分析を行う．

1. 都道府県ごとのツイート数（件数順）
2. Number of Tweets by Prefecture

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 東京都 | 神奈川県 | 大阪府 | … | 徳島県 |
| 6,429 | 2,604 | 2,294 | … | 106 |

## 二値化とデータ研磨の閾値

適切な閾値を設定するために行う．

まず，計算した各辺のNPMIに対し，閾値を0.1から0.9まで0.1刻みで設定し，二値化を行う．二値化による閾値ごとの辺数の変化を表2に示す．表2より閾値の値を大きくすると辺数が減少し，関係性の強い単語同士の辺が残っていることが分かる．今回は，単語同士の関係性が高い0.6から0.9の範囲から，最も辺数が減少したことで関係性の強い辺だけが残った0.6を二値化の閾値とする．

1. 二値化後の辺数
2. The number of edges for each threshold

|  |  |  |  |
| --- | --- | --- | --- |
| 二値化の閾値 | 辺数 | 二値化の閾値 | 辺数 |
| 0.1 | 77,243 | 0.6 | 6,425 |
| 0.2 | 62,659 | 0.7 | 1,486 |
| 0.3 | 45,809 | 0.8 | 881 |
| 0.4 | 30,274 | 0.9 | 401 |
| 0.5 | 17,203 |  |  |

次に，データ研磨を行う際の閾値を決める．二値化の閾値同様に，閾値が小さすぎるとグラフが膨大な辺数になってしまうため，0.6から0.9までを閾値の範囲とした．閾値ごとの辺数と極大クリークの数の変化を表3に示す．表3より，0.7の閾値の時に辺数が大きく減少していることが分かる．さらに，クリークの数も多いため，多くのトレンドを抽出できると考えた．そのため，今回は，0.7をデータ研磨の閾値とする．

1. データ研磨後の辺数，極大クリーク数
2. The number of edges and maximal cliques after data polish for each threshold

|  |  |  |
| --- | --- | --- |
| データ研磨の閾値 | 辺数 | 極大クリーク数 |
| 0.6 | 11,648 | 330 |
| 0.7 | 5,419 | 405 |
| 0.8 | 2,619 | 189 |
| 0.9 | 1,199 | 35 |

## 都道府県ごとの特徴的なトレンド

都道府県の特徴的なトレンドを抽出するために行う．

まず，データ研磨を行ったグラフの極大クリークの一部を表4に示す．表4の単語集合を調べると11月30日の前後で開催されたイベントや行われる前のイベントであることが確認できたため，トレンドを抽出できたと言える．

1. 東京都のツイートから抽出したクリークの例
2. Examples of extracted clique from tweets in Tokyo

|  |
| --- |
| 鏡，目白，ライトアップ，X，モミジ，庭園，照らす，秋 |
| 表参道，()，イルミネーション，よる，並ぶ |
| 20時，連，間，引ける，ミッション，分，シノアリス，配布，チケット，リニューアル，本日，上位，日，4位，22時，トレンド |

次に，東京都の特徴的なトレンドを抽出するために東京都以外のツイートからも，トレンドを抽出する．極大クリークを列挙したもの一部を表5に示す．

1. 東京都以外のツイートから抽出したクリークの例
2. Examples of extracted clique from tweets other than Tokyo

|  |
| --- |
| 横浜，散歩，みなとみらい，夜景，海，休日，空，Yo，キレイ，イルミネーション |
| 炎の体育会TV，困り顔，体育会，炎，対抗，松本人志，チャレンジ" |
| 20時，連，間，引ける，ミッション，分，シノアリス，配布，チケット，リニューアル，本日，上位，日，4位，22時，トレンド |

最後に，東京都のトレンドと東京都以外のトレンドの比較を行う．東京都の極大クリークと東京都以外の極大クリークを比較した結果，東京都のトレンド405件の内，10件が全国的なトレンドとして除外された．除外された極大クリークの一部を表6に示す．以上により，都道府県ごとの特徴的なトレンドが抽出できたと言える．

1. 除外されたクリークの例
2. Examples of excluded clique

|  |
| --- |
| Lupin，当たる，ステップ，スペシャル，チャンス，企画，有馬記念，盗む，五，ギフト，セット，電子マネー，ルパン三世，クリア，黄金，有馬，馬券，1万円，64枚，シ |
| 20時，連，間，引ける，ミッション，分，シノアリス，配布，チケット，リニューアル，本日，上位，日，4位，22時，トレンド |

# むすび

本研究では，収集した全国の位置情報付きツイートを用いることで，都道府県ごとの特徴的なトレンドを抽出することができた．今後の課題として，トレンドの抽出を継続して行いトレンドの動向を探ることでより信憑性のあるトレンドが解析できると考えられる．

1. [We Love Social](https://blog.comnico.jp/we-love-social/)，【2019年11月更新】人気SNSの国内＆世界のユーザー数まとめ（Facebook、Twitter、Instagram、LINE）, https://blog.comnico.jp/we-love-social，2019年12月16日参照.
2. 宇野他, データ研磨によるクリーク列挙クラスタリング, 情報処理学会研究告，Vol. 2014-AL-146，No. 2，pp. 1-8, 2014.