1H2-1

位置情報付きツイートを用いたユーザ属性推定と地域クラスタリング

Presuming the Attribute of Twitter Users and Area Clustering by Analyzing geo-tagged Tweets

近藤 聖也[*1] 吉田 孝志[*2] 和泉 潔[*1] 山田 健太[*1]
Seiya Kondo Takashi Yoshida Kiyoshi Izumi Kenta Yamada

[*1] 東京大学大学院 工学系研究科
School of Engineering, The University of Tokyo

[*2] 日本電気株式会社 情報・ナレッジ研究所
Knowledge Discovery Research Laboratories, NEC Corporation

By the popularization of smart phone and social networking service, many users generate information of their activities and feelings with their locational information. We believe these high volume and high variety data sets enable us to understand their activities in a specific area. We propose a method to presume the attribute of users by analyzing the contents of geo-tagged tweets. Using presumed user attribute and locational information, we perform clustering of a specific area and grasp its features.

1. はじめに

近年，スマートフォンやソーシャルネットワーキングサービス(SNS)の普及により，多くのユーザが自らの活動や感情を位置情報とともに発信するようになった．このように発信されたテキスト情報を分析して，マーケティングに応用するサービスに注目が集まっている．さらに，マーケティングにおいて，地域ごとの特性を考慮した，エリア・マーケティング戦略が求められている．

[李 2012]では，位置情報付きツイートの位置情報データから，特定の地域におけるツイート数やユーザ数，移動ユーザ数を算出し，地域の特徴を抽出している．しかし，ツイッターの特性上，ユーザの性別や年齢といった属性が不明であることから，ユーザの属性を考慮した地域の特徴については言及されていない．

ツイッターのユーザの属性に関する研究の一つとして，[池田 2012]では，ツイートの内容を分析することで，ユーザの性別や年代，居住地を推定する手法を提案している．また，このような基本属性だけではなく，さらに深い属性として，[Preotiuc-Pietro 2015]では，ユーザの年収を推定する手法を提案している．このような深い属性を推定したことは，ツイッターのデータならではの研究結果であると言える．このようなユーザの深い属性は，位置情報と結びつけることで，エリア・マーケティング戦略立案に活かすことが出来るが，位置情報とツイートの内容の両方を用いて分析を行った研究はまだ数少ない．

そこで本研究では，位置情報とツイートの内容の両方のデータを用いて，地域の特徴付けを行った．はじめにツイートの内容を分析することで，ユーザの属性を推定するモデルを構築した．具体的には，テキストを形態素解析し，ユーザごとに特徴的な単語の出現回数行列を作成することで，属性を推定する分類器を作成した．次に，このモデルにより推定したユーザ属性をもとに，地域の特徴を抽出し，クラスタリングを行った．このようにユーザの属性を使った地域クラスタリングを行うことで，細かいエリアごとのマーケティング戦略の立案に活かすことが出来る．

本論文の構成は以下のとおりである．2 章では位置情報付きツイートデータを用いたユーザ属性の推定手順と，その結果について説明する．3 章では 2 章の結果を用いて，地域の特徴抽出とクラスタリング結果を述べる．4 章で本研究のまとめと，今後の展望について述べる．

連絡先：近藤 聖也，東京大学大学院工学系研究科，
kon.arsenal@gmail.com

2. ユーザ属性の推定

本節では，ユーザ属性の推定手順と，その結果について述べる．本研究ではユーザ属性として，具体的には性別と年代について推定した．なお，本研究で用いた位置情報付きツイートの期間は，2013 年 1 月 1 日から 2014 年 12 月 31 日までである．

2.1. 性別推定

ツイッターでは，ユーザのプロフィールとして，自己紹介文を記載するフィールドがある．本研究では，まず，正解データを作る必要があるので，ランダムに抽出したユーザの自己紹介文を読み，明らかに性別が断定できるユーザにのみ性別(男性，女性)のラベル付けを行った．その結果，男性，女性各 563 ユーザの正解データを作成した．これらのユーザの期間内の全位置情報付きツイートを取得したところ，全部で 271,637 レコードのツイートを取得した．

次に，ツイートの内容を形態素解析し，単語ごとに出現回数を計算した．そして，男女の属性の特徴を表す単語の抽出を行った．特徴的な単語を抽出することで，計算量を抑え，実用的で簡潔な分類器を作ることが可能になる．単語を抽出する指標として，クラスcにおける単語wの重要度$I_{w,c}$を，次の式(1)のように定義した．

$$I_{w,c} = \frac{\sum_{j=1}^{n_c} \frac{tweet\_count_{j,w}}{tweet\_count_j}}{\sum_{i=1}^{n} \frac{tweet\_count_{i,w}}{tweet\_count_i}} \quad (1)$$

ここで，nは全体のユーザ数，n_cはクラスcに属するユーザ数，$tweet_count_i$はユーザiのツイート数，$tweet_count_{i,w}$はユーザiの単語wを含むツイート数をそれぞれ表している．

属性に関係なく一般的に用いられる単語は，$I_{w,c}$が 1 に近づき，特定の属性のユーザが，平均的によく投稿する単語ほど，$I_{w,c}$は大きな値を取ることになる．この重要度が高い上位 100 単語をその属性の特徴単語として用いた．しかし，このままだと，固有名詞のように，特定のユーザだけが頻繁に投稿する特殊な単語が上位に来ることがあり，これは全体的な男女の傾向を表すものではないので，全体で 100 回以上出現した単語に絞った上で，重要度の高い単語を抽出した．その結果，男性クラスの重要度が高い単語には，「俺」，「オレ」，「おっさん」等が含まれ

ていた．女性クラスの重要度が高い単語には，「わたし」「あたし」「♡」等が含まれていた．

ユーザごとに，男性，女性クラスの重要度上位 100 単語（計 200 単語）の出現回数行列を作成した．しかし，このまま出現回数で分類器を構築してしまうと，ユーザごとのツイートの回数の差による偏りが生じてしまうため，各ユーザのツイート回数で単語の出現回数を割り，各ユーザの1ツイート当たりの単語出現回数行列を作成し，これを入力として分類器を作成した．本研究では決定木分析を用いて，モデル化を行った．男性，女性それぞれ 563 サンプルのうち，各 463 サンプル，計 926 サンプルで分類器を作成し，これを残り 200 サンプルをテストデータとして当てはめ，モデルの精度を評価した．テストデータに当てはめた際の混同行列は次の表 1 のようになった．

表 1　性別推定の混同行列

		実際のクラス	
		男性	女性
予測クラス	男性	63	11
	女性	37	89

表 1 を見てみると，全体の正解率は 76%となった．特に女性の正答率は約 90%であるが，一方で男性を女性と誤判別することが多い結果になった．これは，本研究で作成した決定木モデルにおいて，分類に用いた 200 単語が一度も出現していないユーザは，女性と判別されることが原因として考えられる．

2.2. 年代推定

2.1 節では性別推定手法について説明したが，本節では同様の手順で年代推定を行った結果について述べる．まず自己紹介文を読み，年代についての正解データを作成した．この際，年代を，[池田　2012]と同様に 10 代，20 代，30 代，40 代以上，という 4 つのクラスに分類した．その結果，各クラス 319 ユーザの正解データを作成できた．これらのユーザの期間内の全ツイートを抽出したところ，262,317 レコード抽出できた．このデータに対し，2.1 節と同様の手順で，特徴単語の抽出と，モデル構築を行った．

特徴単語を抽出する際，男女の判別と同様に，全体で 100 回以上出現した単語だけに絞り，クラスごとに式(1)の重要度が高い上位 100 単語を抽出した．その結果，10 代には「体育」，「部活」等が，20 代には「若い」，「早起き」等が，30 代は「ゴルフ」，「ビール」等が，40 代は「週末」，「連休」等が特徴単語として抽出された．

これらの計 400 単語に対して，ユーザごとに 1 ツイート当たりの単語の出現回数行列を作成し，決定木分析を行った．各クラス 239 サンプルずつを学習データとして用い，残り 80 サンプルずつをテストデータとしてモデルを当てはめ，次の表 2 のような結果が得られた．

表 2　年代推定の混同行列

		実際のクラス			
		10 代	20 代	30 代	40 代～
予測クラス	10 代	59	43	22	19
	20 代	10	22	5	5
	30 代	4	5	42	6
	40 代～	7	10	11	47

結果として，全体の正解率は 53%程度となった．表 2 を見てみると，特に 20 代を 10 代と誤判別することが多く見られ，10 代と 20 代の判別が困難であった．これは，10 代の特徴単語として挙げられた，「部活」等の学生を表すような単語を，20 代の学生ユーザも使用するためだと考えられる．

3.　地域クラスタリング

本章では，2 章で構築したユーザ推定モデルを用いて，地域の特徴付けを行う手法について説明する．本研究において対象とした地域はお台場地域である．お台場地域を図 1 に示すように，施設ごとに 15 個の小エリアに分割し，この小エリアに対して地域クラスタリングを行った．お台場は，ショッピングモールや，駅，学術施設，イベントエリア等，対象とするユーザ層の異なる様々な施設が立ち並んでいるため，本研究のように，ユーザ属性を用いた地域クラスタリングの対象として，適したエリアであると考えた．

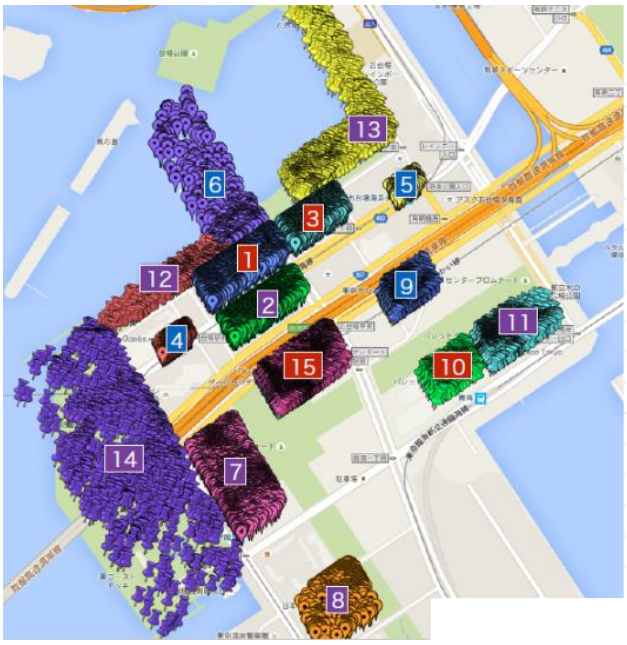

図 1　お台場地域のエリア分け

地域クラスタリングを行う手順として，まず 2 章で得られた分類モデルを当てはめ，お台場エリアでツイートしたユーザの属性を推定する．ここで，モデルを当てはめる際に用いたツイートデータは，2013 年 1 月 1 日から 2014 年 12 月 31 日までの期間に，お台場エリア内において，1 回以上ツイートしたことのあるユーザの全ツイートデータである．合計で 18,035 ユーザの 1,467,913 ツイートが得られた．

次に，15 個の小エリアそれぞれに対し，時間帯，曜日ごとに，ツイートをしたユーザの比率を求める．具体的には，時間帯は「0 時から 6 時まで」，「6 時から 12 時まで」，「12 時から 18 時まで」，「18 時から 24 時まで」の 4 分類を用いた．曜日は月曜から金曜の「平日」，土曜と日曜の「休日」の 2 分類を用いた．この中で，「10 代男性」，「20 代女性」といった年代 4 分類×性別 2 分類の計 8 分類のツイート数の相対比率を計算した．従って，合計で 4×2×8＝64 次元の特徴量を計算し，これを各エリアで算出して，k-means によるクラスタリングを行った．なお，各特徴量は k-means に当てはめるために，平均 0，標準偏差 1 となるように標準化した．

このようにして得られた特徴量を用いて，k-means によるクラスタリングを行うことで，似た特徴を持つ地域をグルーピングした．なお今回は 3 クラスに分類を行った．クラスタリングの結果は次の表 3 のようである．

表 3 地域クラスタリングの結果

番号	施設名	分類番号
1	Aqua City	1
2	フジテレビ	2
3	DECKS	2
4	台場駅	0
5	お台場海浜公園駅	0
6	水上バス	0
7	イベントエリア	1
8	日本科学未来館	1
9	東京テレポート駅	2
10	パレットタウン	0
11	観覧車	2
12	自由の女神	0
13	砂浜	2
14	都立潮風公園	1
15	Diver City	2

この結果を解釈すると，0 番に分類されたのは，駅や水上バス等の交通機関が多い．1 番に分類されたのは，日本科学未来館やイベントエリア等のイベントが行われる施設が多く，2 番に分類されたのは，フジテレビや DECKS 等の娯楽施設が多い．このように，各施設の持つ特徴ごとに，そこに存在するユーザの属性を用いて，分類をすることが出来た．

4. まとめ

本研究では，位置情報付きツイートデータを用いて，まずユーザの属性を推定する手法を提案した．結果として，性別については 76%，4 つに分類した年代については 53%程度の精度で分類することが出来た．次に，このユーザ属性を推定するモデルを用いて，お台場の各施設のクラスタリングを行った．その結果，交通機関，イベントがあるエリア，娯楽施設の 3 つ施設に分類を行うことが出来た．

今後の展望として，ツイッターの特徴である，ユーザの感情や詳細な活動を分析することで，居住地，趣味，職業，年収等の，より深い属性を調べることが挙げられる．また，今回行ったような小さいエリア分けでは，周囲のエリアとの物理的距離の近さにより，特徴が出にくい可能性があるので，さらに広いエリアでの地域のクラスタリングを行う．

参考文献

[池田 2012] 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 マーケット・デバイス＆システム，Vol.2，No.1，82-93，(2012).

[李 2012] 李 龍，若宮 翔子，角谷 和俊: Tweet 分析による群衆行動を用いた地域特徴抽出，情報処理学会論文誌データベース，Vol.5 ，No.2，36-52，(2012).

[Preotiuc-Pietro 2015] Preotiuc-Pietro D，Volkova S，Lampos V，Bachrach Y and Aletras N:Studying User Income through Language，Behavior and Affect in Social Media，PLOS ONE (2015).