4M1-4

Twitter 本文を用いた観光情報抽出及び分析システムの構築

Extraction of Tourist Information from Contents of Tweets and Building an Analysis System

小原 基季[*1] Motoki OHARA　森田 和宏[*1] Kazuhiro MORITA　泓田 正雄[*1] Masao FUKETA　青江 順一[*1] Jun-ichi AOE

[*1] 徳島大学大学院先端技術科学教育部
Graduate School of Advanced Technology and Science, The University of Tokushima

Twitter is possible to transmit information with the position information called geotag. This feature has been used to sutdy the extraction of behavior analysis and tourist information. However, geotagged tweets are only a few percent of the total.In this study, we propose extraction methods of tourist imformation from contents of tweets. The purpose of this paer is to obtain tourist information from tweets with or without geotag.Also, we built an analysis system of the extracted tourist information.

1. はじめに

平成 21 年，観光庁により，「観光入込客統計に関する共通基準」（以下「共通基準」）が策定された．従来，観光に関する統計は各都道府県によって手法が異なっていたため，地域間での比較が困難だった．そこで，各都道府県の観光統計を整備するため，共通の把握方法として共通基準が制定された[観光庁 2013]．

観光入込客調査の実施は，都道府県と市区町村の担当者が観光地点での入込客調査の実施や結果の整理などをおこなうため，人手の確保が必要となる．また，観光客に対してアンケート調査を実施するため，大変なコストがかかってしまうこと，結果を集計し，公表するまでに時間がかかってしまうという問題点がある．そこで，個人が情報を容易に発信でき，データが無償で公開されている Twitter を用いて観光情報を取得し，分析をおこなう．

マイクロブログの 1 種である Twitter は，Tweet と呼ばれる 140 文字以内のメッセージをパソコンや携帯端末を用いて投稿できるサービスである．Tweet には，ジオタグと呼ばれる位置情報を付与して情報を発信することができ，ユーザーは外出先から位置情報付で Tweet することで，どこで何をしているのかを周囲に知らせることができる．このジオタグ機能を用いてユーザーの行動分析，観光情報の抽出をおこなう研究[酒巻　2012][桐村 2013][中嶋　2013]が盛んにおこなわれている．しかし，ジオタグ付きの Tweet は全体の数パーセントに過ぎず，一部の Tweet にしか対応できていないという問題点がある．解決策として，Tweet の本文から地域連想語[晃昇　2012]や観光に関する Tweet に用いられやすい単語（なう，到着した，楽しかった等）を用いて観光情報を抽出する方法が考えられる．本研究では，Tweet 本文を用いることで，ジオタグの有無に関わらず観光情報を抽出することを目的とする．また，抽出した観光情報を分析するシステムを構築する．

2. 関連研究

Twitter を用いてユーザーの行動分析をおこなう研究として酒巻は，ジオタグ付き Tweet の位置情報，時刻情報，投稿内容を用いてクラスタリング，ラベリングをおこない，「定期的にどこで活動しているか」「その場所でどのような活動をしているのか」について解析をおこなった[酒巻　2012]．

ユーザーの観光行動を分析する研究として桐村は，ジオタグが付与された Twitter の投稿データを利用して，ユーザーの行動の基本的な特徴を把握し，観光行動の分析例を示した．分析結果から，ユーザーの日常的な生活圏は二大都市圏にやや偏っているものの，その行動範囲は週末になると広くなる傾向が確認でき，観光等の余暇活動が Tweet に現れていることを示した[桐村　2013]．中嶋らは，観光名所付近でつぶやかれたジオタグ付き Tweet を検索し，旅行者の Tweet に現れる特徴や，Instagram 等のサービスから観光名所毎に観光ツイートを収集し，旅行者のタイムラインから観光ルートを抽出した．また，収集した観光ツイートを「食事」，「景観」，「行動」に分類し，旅行者の好みに合わせた観光ルートを推薦する手法を提案した[中嶋 2013]．

これらの研究では，Tweet を抽出する際に，ジオタグに大きく依存しており，ジオタグが付与されていない Tweet に対応していないという問題点がある．

本研究では，Tweet の本文を用いることで，ジオタグの有無に関わらず観光情報の抽出をおこなう．

3. 提案手法

本研究では，Tweet から地域連想語，パターンマッチングを用いて，観光情報の抽出，ユーザーの居住地の把握をおこなう．また，抽出した観光情報を用いて分析システムの構築をおこなう．

地域連想語とは，地名や特産品，施設名のように特定の都道府県を連想することができる単語のことを指す．例えば，徳島県の連想語だと“阿波踊り”や“徳島市”が挙げられる．

パターンマッチングでは，単語の表記や品詞等を概念化し，照合規則として用いている．

3.1 観光情報の取得

Step1. Tweet の取得

Twitter API を用いて，Tweet を取得する．この際，リツイートや「Foursquare」，「今ココなう！」等のアプリを用いた Tweet は除去する．

Step2. 形態素解析

取得した Tweet に対して形態素解析をおこない，表記や品詞情報等を取得する．

連絡先：小原基季，徳島大学大学院先端技術科学教育部システム創生工学専攻知能情報システム工学コース，〒770-8506，徳島市南常三島町 2-1，E-MAIL：c501437010@tokushima-u.ac.jp

表 1. 観光情報抽出手法で使用した照合規則例

照合規則	Tweet 例
＜地名＞＜なう＞	徳島なう
＜地名＞＜行動＞	阿波踊り行ってきまーす！
＜地名＞＜感想＞	徳島ラーメン美味かった
＜地名＞＜存在＞	阿波踊り！

＜地名＞= 地域連想語
＜ なう ＞= なう, なぅ, なーう, なう～
＜行動＞= 行っ, 行く, 見てる, 到着, 食べ
＜感想＞= 楽しかった, 美味し, 最高, きれい
＜存在＞= 記号, 助動詞「だ」, いました

Step3. 地域連想語の取得
地域連想語辞書を用いて Tweet に含まれる地域連想語を取得する.

Step4. パターンマッチングを用いた観光情報の抽出
Step2 で取得した形態素解析結果と Step3 で取得した地域連想語を用いてパターンマッチングをおこない, 観光に関する Tweet を取得する. 照合規則の一部を表 1 に示す.

3.2 居住地の推定

Twitter のプロフィール欄には,「場所」という欄があり居住地を記述できるようになっている. しかし, 空欄になっているユーザーや記述されていても地名の記述方法が様々で, 架空の地名や複数の地名を記述しているユーザーが多数存在している.

よって本研究では, Tweet の本文を用いて居住地の推定をおこなう.

Step1. Tweet の取得
Twitter API を用いて対象ユーザーの Tweet を 200 件取得する. この際, リツイート等の不要な Tweet を除去する.

Step2. 形態素解析
取得した Tweet に対して形態素解析をおこない, 表記や品詞等の情報を取得する.

Step3. パターンマッチングを用いた居住地以外の Tweet 除去
Step2 で取得した形態素解析結果を用いてパターンマッチングをおこない, 居住地以外の地域で投稿された Tweet を特定し, 除外する. 照合規則の一部を表 2 に示す.

Step4. 居住地の取得
残った Tweet 中から地域連想語を参照し, 各 Tweet を“地域なし”と 47 都道府県の 48 種類に分類する. 分類された各都道府県の内, 分類数が最多の地域をユーザーの居住地とする.

3.3 分析システムの構築

抽出した観光情報, 居住地情報を用いて, 分析システムの構築をおこなう. 出力内容は, Tweet に含まれる地域連想語の割合, Tweetしたユーザー数の推移, Tweetされた場所の分布, ユーザーが住んでいる地域の分布の 4 つとする.

- **Tweet に含まれる地域連想語の割合**
  抽出した観光情報から地域連想語を取得し, 円グラフを用いて割合を表示する. 表示する内容は, 全ての地域連想語を用いたものと地域連想語の中で地名を除いたものの 2 つとする.
- **Tweet したユーザー数の推移**
  観光情報を Tweet したユーザー数の日毎の推移を棒グラフを用いて表示する.

表 2. 居住地推定手法で使用した照合規則例

照合規則	Tweet 例
＜地名＞＜から＞＜帰る＞	明日, 神戸から帰ってきます！
＜地名＞＜土産＞	まさとしくんから沖縄土産届きました
＜地名＞＜うろうろ＞	梅田うろうろしてる
＜実家＞＜地名＞	今嫁の実家の埼玉着いた

＜から＞ = 格助詞「から」
＜帰る＞ = 帰, 帰り道, 帰路
＜土産＞ = 土産
＜うろうろ＞= うろうろ, ぶらぶら
＜実家＞ = 実家

- **Tweet された場所の分布**
  都市名, 施設名から Tweet された場所を求めるため, 次の処理をおこなった. 初めに, 取得した地域連想語を Geocoding API を用いて緯度, 経度に変換する. 次に, 取得した緯度, 経度を用いて市区町村名を取得する. これにより, 異なる地域連想語でも同じ市区町村内に存在するものを 1 つにまとめることができる. 最後に, 市区町村名から緯度, 経度を取得し, Google Maps API より地図上に分布を表示する.
- **ユーザーが住んでいる地域の分布**
  居住地推定手法より取得した居住地情報を Google Maps API を用いて地図上に分布を表示する.

4. 評価実験

観光情報抽出手法, ユーザーの居住地推定手法の有効性を確認するため精度実験をおこなった.

4.1 観光情報抽出手法の精度実験

観光情報抽出手法の有効性を確認するため, 2014 年 7 月 19 日から 2014 年 8 月 11 日までの徳島県内での Tweet を対象として精度実験をおこなった. 人手で正誤判定をおこない, 適合率を用いて評価をおこなった.

結果として, 502 件の Tweet を抽出することができた. また抽出した Tweet の内, 正解件数は 403 件, 適合率は 80.2%と良好な結果が得られた. しかし, 他人の行動に関する Tweet や天気・災害に関する Tweet を誤って取得してしまっていた. 今後, 照合規則の拡充によって誤抽出を減らしていくことを考えている.

4.2 居住地推定手法の精度実験

居住地推定手法の有効性を確認するため, プロフィール欄に居住地について記載のあるユーザー38 人を対象として精度実験をおこなった. 人手で正誤判定をおこない, 正解率を用いて評価をおこなった.

実験結果は, 正解数 31 人, 正解率は 81.6%となった. 尚, 正解地域が他の地域と同数で 1 位となった場合は準正解として, 正解数に含めた.

地域連想語を含む Tweet の数には個人差が大きく, Tweet 中に地域連想語を数回, あるいは一度も含まないユーザーが存在した. また, 人名を地域連想語として誤検出しているものがいくつか見られた.

今回の手法では, 居住地以外の Tweet をパターンマッチングを用いて除去しているが,「徳島に住んでいる」のような居住地と特定できる照合規則を作成すれば, Tweet 中に地域連想語を

含む割合の少ないユーザーにも対応できるのではないかと考えている.

5. 分析システム

抽出した観光情報, 居住地情報を用いて分析システムの構築をおこなった. 徳島県と北海道についての出力結果を以下に示す.

5.1 徳島県での出力結果

Twitter API より徳島県全域が含まれるように緯度・経度, 半径を検索に用いて得られた 2014 年 8 月 12 日から 2014 年 8 月 15 日までの観光情報に関する Tweet を対象とした. 地域連想語の割合, 観光客の推移のグラフを図 1 から図 3 に示す.

解析結果から地域連想語の割合について,「阿波踊り」や演舞場のある「藍場浜」,「南内町」について多くの割合を占めていることが分かった. これは, 対象が阿波踊り期間中だったためだと考えられる. また, 観光客の推移について初日が一番少ない結果となった.

5.2 北海道での出力結果

北海道での分析には, 2015 年 3 月 3 日から 2015 年 3 月 9 日までの Tweet を対象とし, ジオタグを利用して収集した Tweet とジオタグを利用せずに収集した Tweet の 2 つのデータを用いた. 1 つ目は, Twitter API より北海道全域が含まれるように緯度・経度, 半径を検索に用いて得られた観光情報に関する Tweet, 2 つ目は, 1 月 1 日から 1 月 7 日までの観光情報分析結果から 10 件以上 Tweet された地域連想語を検索に用いて, 得られた観光情報に関する Tweet を対象とした.

出力結果から, 地域連想語を検索に用いて収集したデータの方が,「北海道」を含む割合が大きかった. これは, 道内にいる人は「北海道」と Tweet するよりも「札幌」や「函館」のような都市名を Tweet する場合が多いからではないかと考えられる. また,「ジンギスカン」や「札幌ラーメン」のような食べ物に関する割合も地域連想語を検索に用いたものの方が多くの割合を占めていることが分かった. これは道外でこれらの食べ物を食べた人の Tweet を取得してしまうからだと考えられる.

観光客の推移について見ると, 緯度・経度, 半径を検索に用いて取得したものについては, 最少は 3 月 3 日の 85 人, 最多は 3 月 8 日の 128 人なのに対して, 地域連想語を検索に用いて取得したものは, 最少で 3 月 3 日の 7,152 人, 最多で 3 月 8 日の 9,328 人となった. 人数の推移について見ると両方のグラフに差があまり見られなかったが, 人数については地域連想語を検索に用いた方が大きく上回ることが分かった.

6. まとめと今後の課題

本稿では, Twitter 本文を用いた観光情報の抽出手法, ユーザーの居住地推定手法, 観光情報を用いた分析システムの構築について述べた. また, 観光情報抽出手法についての精度実験, 居住地推定手法についての精度実験, 分析システムの動作についての考察をおこなった.

今後の課題として, 照合規則の拡充, 実際の観光データとの比較をおこなうことが必要である.

参考文献

[観光庁 2013] 国土交通省観光庁: 観光入込統計に関する共通基準, 2013.

[酒巻 2012] 酒巻智宏: マイクロブログのジオタグを用いたユーザの行動分析, 東京大学大学院修士論文, 2012.

[桐村 2013] 桐村喬: 位置情報付きツイッター投稿データにみるユーザー行動の基本特徴―観光行動分析への利用可能性―, 地理情報システム学会講演論文集 22, 2013.

[中嶋 2013] 中嶋勇人, 新妻弘崇, 太田学: 位置情報付きツイートを利用した観光ルート推薦, 情報処理学会研究報告, pp.1-6, 2013.

[晃昇 2012] 晃昇祥恵: 地域連想語辞書の構築に関する研究, 言語処理学会第 18 回年次大会, pp.1095-1097, 2012.

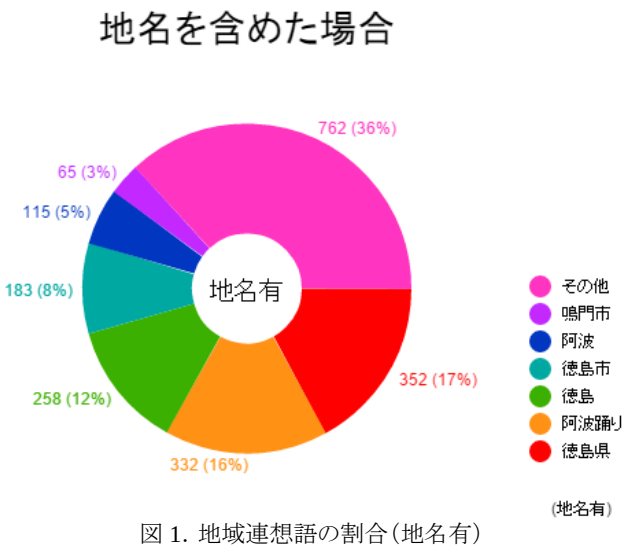

図 1. 地域連想語の割合(地名有)

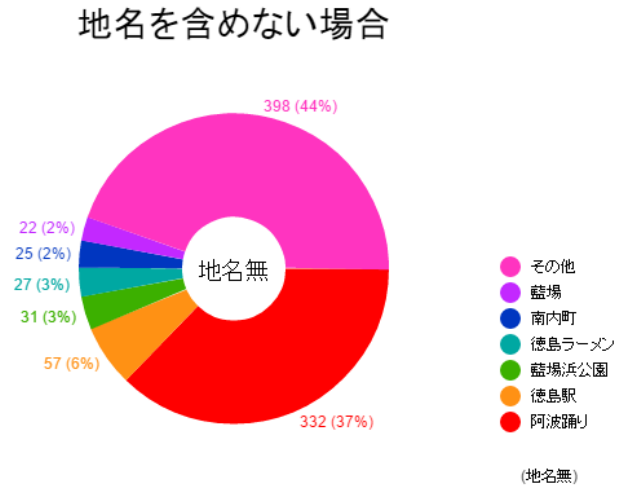

図 2. 地域連想語の割合(地名無)

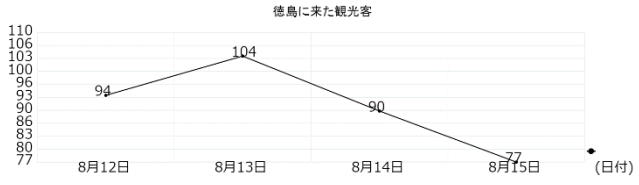

図 3. 観光客の推移