G-15

位置情報付きツイートと観光地名入りツイートを用いた位置情報無しツイートからの観光情報抽出手法の提案

Proposal of Tourist Information Extraction Methods from Tweets without Position Information by Tweets with Position Information and Tweets Containing Tourist Spots Names

渡邉　小百合 †　　吉野　孝 †
Sayuri Watanabe　Takashi Yoshino

1. はじめに

観光庁による旅行・観光消費動向調査 [1] では，2006 年から減少を続けていた日本人の旅行平均回数が，2010 年以降からその減少が止まっている．これは，ドラマやアニメの舞台への聖地巡礼等の新しい形態の旅行が出てきたことが要因として考えられる．また，観光庁による訪日外国人旅行者数・出国日本人人数の推移 [2] では，外国人旅行者が 2012 年から年々増加している．これより，ドラマ・アニメとのコラボイベントや外国語への対応といった観光地に対して新しいニーズが発生することが考えられるため，観光地側も観光客のニーズや問題点を知り，観光地の発展につなげていく必要がある．

先行研究では，Web 上から観光情報を抽出し，類似性を可視化するシステムの開発を行ったが，Twitter から観光情報があまり得られないという問題があった [3]．その原因として,「観光地名が入っていない観光地に関するツイート」を考慮していなかったことが挙げられる．たとえば,「偕楽園の梅見は，お花にはややがっかり感がありました」のような，ツイート内に観光地名が含まれているものは収集出来るが,「今年の梅の花は微妙」のような，観光地名を含んでいない観光地に関するツイートは収集出来ていない．

そこで，本研究では，位置情報付きツイートと観光地名入りツイートの特徴語を用いた観光情報抽出手法を提案する．観光地周辺の位置情報が付加されたツイートは，観光地名が入っていなくても観光地に関する情報である可能性が高い．しかし，位置情報が無く，観光地名も含まれていないツイートから観光情報を発見するのは困難である．これより，観光地名の入ったツイートと観光地周辺の位置情報付きツイートの特徴語を用いて，位置情報が無く観光地名も入っていないツイートから観光情報を抽出する．

2. 関連研究

奥らは，位置情報付きツイートと位置情報付き写真を用いた観光スポット推薦システムを開発した [4]．対象とする観光地名を含んだ位置情報付きツイートによる観光地の活動領域と，位置情報付き写真による観光地の活動領域を合成して，観光地の活動領域を推定し，その領域内に含まれるツイートの特徴から観光スポットの推薦を行っている．Lee らは，Twitter から社会的イベントを検出するための，ツイートの時間と位置情報を用いた地理的規則性の測定手法を提案した [5]．対象領域において，位置情報付きツイートの呟かれた時間，位置，ユーザの行動から通常の地理的規則性を推定し，その規則性から外れている時をイベントとして検出する．

どちらの研究においても，位置情報付きツイートを用いて対象領域を推定する点とその特徴を利用する点では同じである．しかし，本研究では位置情報付きツイートと観光地名入りツイートを用いて，位置情報が無く観光地名も入っていないツイートから観光地に関する情報の抽出を行う．

† 和歌山大学システム工学部，Faculty of Systems Engineering, Wakayama University

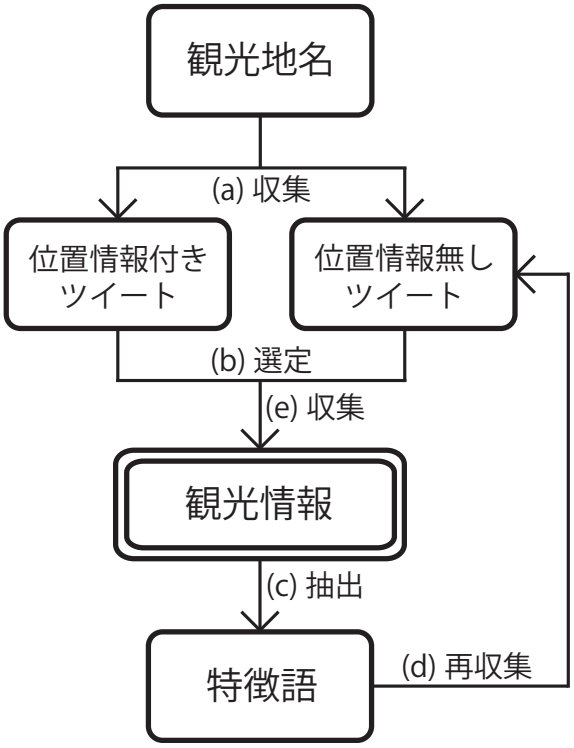

図 1: 観光情報抽出手順

3. 観光情報抽出手法

本研究では，位置情報付きツイートと観光地名入りツイートの特徴から，位置情報が無く観光地名も入っていないツイートから観光情報を抽出する．図 1 に観光情報抽出手法の手順を示し，以下にその具体的な内容を示す．

(1) 観光地名入りツイートの収集

位置情報付きツイートと位置情報無しツイートから，観光地名を含むツイートを収集する (図 1(a))．位置情報付きツイートにおいては，奥らの活動領域推定手法 [4] を用いて，観光地に関することがツイートされる領域を推定し，その範囲内のツイートも収集する．収集したツイートの中から観光情報になりうるものを選定する (図 1(b))．その選定は，現在は人手で行っているが，今後自動化する予定である．

(2) 特徴語の抽出

収集された観光情報を含むツイートの特徴語を抽出す

表 1: 「伏見稲荷」で収集したツイート数と観光情報の数

ツイートの種類	収集したツイート数	観光情報数
位置情報付きツイート	7314	20
位置情報無しツイート	1605	180

表 2: 「伏見稲荷」で収集したツイートの特徴語の例

特徴語	tf-idf 値
京都	0.479
千本鳥居	0.422
行って	0.217
写真	0.159
おもかる石	0.080

る (図 1(c)). 形態素解析システム JUMAN[*1] を用いてツイートの分かち書きを行い，特徴語の抽出には tf-idf を用いる．抽出対象の品詞は，観光地の特徴を表す語となりうると考えられる名詞，形容詞，動詞としている．本研究では，tf-idf 値が高い上位 20 件を，再収集を行う特徴語の対象とする．

(3) 位置情報無しツイートの再収集

位置情報無しツイートから，(2) で抽出した特徴語を含むツイートを再収集する (図 1(d)). この収集では，すでに収集したツイートは除外する．

(4) 観光情報の収集

(2) と (3) を繰り返すことにより，位置情報が無く観光地名が入っていないツイートから観光情報を収集する (図 1(e)).

4. 実験

4.1 実験概要

本実験は，提案した観光情報抽出手法が有用であるかを検証する．本実験では，トリップアドバイザーによる日本観光ランキング[*2] において 1 位である伏見稲荷大社についての観光情報を抽出する．本実験で使用するツイートは，Twitter Streaming API を用いて Twitter 上からリアルタイムに収集した，2016 年 1 月 1 日から 2016 年 6 月 30 日の間に呟かれた約 1 億 3000 万ツイートである．これらのツイートは，Twitter Streaming API を用いて Twitter からリアルタイムに収集したものであるが，期間中に収集の停止や二重収集といった不具合が少し起きたため，一部不完全である．使用するツイートのうち，位置情報付きツイートは約 79 万ツイートであった．本実験では,「伏見稲荷」で収集したツイートの特徴語を抽出し，その特徴語で収集したツイートの特徴語の抽出までを行う．

4.2 実験結果と考察

表 1 に「伏見稲荷」で収集したツイート数と観光情報の数を示す．位置情報付きツイート内の観光情報は 20 件であったのに対し，位置情報無しツイート内の観光情報は 180 件と，9 倍の差があった．この結果より，位置情報無しツイート内には観光情報が多く含まれているので，位置情報無しツイートから観光情報を抽出する必要性があることがわかる．表 2 に「伏見稲荷」で収集したツイートの特徴語の例を示す.「京都」や「行って」というような，場所に関する単語の tf-idf 値が高い傾向にあった．また,「千本鳥居」や「おもかる石」というような，伏見稲荷大社特有のものも特徴語として抽出された．その場所特有なものを特徴語として用いることで，より対象とした観光地に関係のあるツイートが収集できると考えたため,「千本鳥居」と「おもかる石」を含むツイートの再収集を行った．

表 3: 2 種類の特徴語で収集したツイート数と観光情報の数

特徴語	収集したツイート数	観光情報数
千本鳥居	104	27
おもかる石	10	3

表 3 に「千本鳥居」と「おもかる石」で収集したツイート数と観光情報の数を示す．観光地名は含まれていないが特徴語を含んでいる「昨日初めて某所の千本鳥居を見たのですが，圧巻でした」,「おもかる石が軽かったことにびっくりしてる」のような，伏見稲荷大社に対する印象や感想のツイートが新たに 30 ツイート収集できた．この結果から，位置情報付きツイートと観光地名入りツイートの特徴語を用いた観光情報の収集は，有用であると言える．表 4 に「千本鳥居」，表 5 に「おもかる石」で収集したツイートの特徴語の例を示す．どちらの結果においても,「鳥居」や「重い」のような，再収集に用いた語句を表す単語が抽出された．しかし,「伏見稲荷」で収集したツイートの特徴語とは違い，伏見稲荷大社特有のものはなかったため,「千本鳥居」と「おもかる石」で収集したツイートの特徴語を用いた収集では，伏見稲荷大社に関する観光情報だけではなく，他の観光地に関する情報も出てくる可能性がある．これより，特徴語によって収集したツイートの中に他の観光地の情報も混ざっている場合，どこまでを対象とした観光地の特徴語とするのか検討する必要がある．また，今回の実験では伏見稲荷大社のツイートを対象としたが，あまり有名ではないマイナーな観光地においても，特徴語から新たな観光情報が抽出できるかの検証を今後行う必要がある．

5. おわりに

本稿では，位置情報付きツイートと観光地名入りツイートの特徴語を用いた観光情報抽出手法を提案した．観光地名の入ったツイートと観光地周辺の位置情報付きツイートの特徴から，位置情報が無く観光地名も入っていないツイートから観光情報を抽出する．実験では,「伏見稲荷」が含まれたツイートの特徴語として「千本鳥居」や「おもかる石」のような，観光地特有のものが抽出された．その特徴語を用いて再収集した結果，観光地名は含まれていないが特徴語を含んでいる，伏見稲荷大社に対する印象や感想のツイートを収集できたため，位置情報付きツイートと観光地名入りツイートの特徴語を用いた観光情報の収集は，有用であると言える．しかし，再収集したツイートの特徴語は,「鳥居」や「重い」のような，他の観光地の観光情報も出てくる可能性があるものであった．今後の課題としては，特徴語によって収集したツイートの中に他の観光地の情報も混ざっている場合，どこまでを対象とした観光地の特徴語とするかの検討や，マイナーな観光地においても特徴語から新たな観光情報が抽出できるかの検証が挙げられる．

参考文献

[1] 国土交通省 観光庁:「旅行・観光産業の経済効果に関する調査研究」(2014 年版), http://www.mlit.go.jp/common/001136064.pdf (参照 2016-7-21).

[*1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?cmd=read&page=JUMAN

[*2] https://www.tripadvisor.jp/Attractions-g294232-Activities-Japan.html

表 4: 「千本鳥居」で収集したツイートの特徴語の例

特徴語	tf-idf 値
鳥居	0.290
しんどい	0.193
圧巻	0.193

表 5: 「おもかる石」で収集したツイートの特徴語の例

特徴語	tf-idf 値
重い	0.632
暗示	0.316
びっくり	0.316

[2] 国土交通省 観光庁: 訪日外国人旅行者数・出国日本人人数の推移, http://www.mlit.go.jp/kankocho/siryou/toukei/in_out.html (参照 2016-7-21).

[3] 渡邉小百合，吉野孝 : 観光地間の類似性を基にした向上点発見のための観光情報可視化システム,「マルチメディア, 分散，協調とモバイル (DICOMO2016) シンポジウム」, pp.1357–1362(2016).

[4] 奥健太，橋本拓也，上野弘毅，服部文夫 : 位置情報付きツイート対応付けに基づく観光スポット推薦システムの開発，ARG 第 2 回 Web インテリジェンスとインタラクション研究会, pp.7–12(2013).

[5] Ryong Lee，Kazutoshi Sumiya : Measuring Geographical Regularities of Crowd Behaviors for Twitter-based Geo-social Event Detection，Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, pp.1–10(2010).