D-026

マイクロブログを用いた地域別特徴抽出手法の一検討
A study of the regional feature extraction method using a micro-blog

遠藤 雅樹†1 †2　荘司 慶行†2　江原 遥†3　廣田 雅春†4　大野 成義†1　石川 博†2
Masaki Endo　Yoshiyuki Shoji　Yo Ehara　Masaharu Hirota　Sigeyoshi Ohno　Hiroshi Ishikawa

1. はじめに

近年，通信ネットワークの高速化やスマートフォン・タブレットなどのデバイスの普及に伴い，常時 Web にアクセス可能となった．そのため，リアルタイム性を持つ情報の発信が容易になり，人やモノが大量のデータを生成し，日々膨大な情報発信が行われている．その中でも，人が生成するデータでは，ソーシャルネットワーキングサービス(SNS)と称される Web サービスが急速に普及し，注目が集まっている．特に，ユーザ数の多い SNS として，Twitter[1]が挙げられる．Twitter などは，人と人とのつながりをサポートするツールであるが，個人だけでなく企業や政府機関[2]においても SNS を利用した情報発信を開始するなど，社会的ネットワークの構築ツールとしても活用されている．

SNS は，リアルタイムなコミュニケーションツールとして利用され，人々が活動をする中で発生する出来事についての様々な情報を発信し共有している．SNS を通じて発信される情報は，情報発信を行うユーザさえいれば，実世界で発生した出来事について，地域や場所，時間帯を問わない即時性・網羅性を持っている．そのため，現在では，テレビや新聞，Web ニュースなどのメディアを補完するインフラとして利用されている．そのため，SNS 上に発信される大量の情報を分析することで，実世界において発生する事象についての状況把握や解決手法を提供できる可能性がある．しかし，SNS の情報は膨大であり，データの質も不均一である．したがって，SNS から有益な情報を抽出する手法は重要なタスクであり，様々な分野で研究が行われている．

我々は， 21 世紀の成長産業として期待される観光に関する視点から SNS の分析を試みる．現在，旅行時の観光情報を取得するには，ガイドブックの活用だけでなく，Web 検索の利用も一般的となった．Web 検索を利用する場合には，ガイドブックなどから得たキーワード，旅行先の「地名」と「観光」・「グルメ」などを組み合わせたキーワードを利用した検索結果からリンク先を閲覧し，さらに旅行先地域の観光に関するキーワードを取得しながら，検索を繰り返す作業が必要である．そのため，検索コストがかかるだけでなく，検索に不慣れな情報弱者の場合には，必要な情報に到達できない場合もある．また，デジタルデバイドによる問題は，Web 上の情報量の増加に伴い，今後，さらに困難になることが予想される．

そこで，我々は，SNS 上の情報を観光情報の集合知と捉え，SNS の持つ即時性・網羅性を活用することで，旅行先地域のご当地グルメやイベントを自動抽出可能なシステムを検討している．SNS は，実世界で発生する出来事について多くのユーザが様々な意見や感情を基に情報発信を行っており，実世界の状況がほぼリアルタイムに SNS に反映されている．そのため，SNS 上に蓄積されている大量のユーザの情報を分析することで，実世界の状況を把握することが可能になっている．

本稿では，SNS の 1 つである Twitter に着目し，位置情報付きツイートをモニタリングし蓄積した時空間ログを用いて各地域の特徴を抽出する手法を提案する．具体的には，収集した位置情報付きツイートの持つ緯度経度情報から導いた地域と一般的な料理名(例：うどん，そば)やイベント名(例：祭り，踊り)を含む語(本稿では対象語と呼ぶ)によって構成した行列を分析することで，各地域に特徴的な対象語を抽出する．提案手法により抽出した対象語は，地域の特徴を表すため旅行先とした地域のご当地グルメや地域のイベントである可能性が高く，抽出した語を利用することで観光情報を提供する上での一助となる．また，SNS の即時性・網羅性を活用するため，ガイドブックに掲載される定番情報だけでなく，最近の流行の取得も可能である．本稿では，旅行先地域に関する知識がない場合であっても，旅行先の観光情報を取得する際に，食やイベントに関する地域別の特徴抽出を行う手法について提案し，その評価を行う．

本論文の構成は次のとおりである．2 章ではマイクロブログから実世界の分析を行った関連研究について述べる．3 章では，Twitter の分析により地域の特徴抽出を行う手法について提案し，4 章で提案手法の評価実験と結果について述べる．最後に 5 章でまとめと今後の課題を述べる．

2. 関連研究

SNS を用いて人々が発信する即時性・網羅性のある位置情報付きの情報に着目して実世界を捉える研究が行われている．土屋ら[3]は，マイクロブログを用いて鉄道の運行トラブル状況を抽出し，復旧状況および混雑状況を判断することで，公式情報だけではできない意思決定を支援する手法を提案している．Sakaki ら[4]は，位置情報付きツイートをモニタリングすることで，台風や地震などの突発的に発生するイベントをリアルタイムに検出するシステムを開発した．Lee ら[5]は，位置情報付きツイートを用いて地域における群集行動の通常性を推定し，異常判定により地域イベントを検出する手法を提案している．李ら[6]は，位置情報付きツイートを地域ごとに分析することで群集行動から特徴的な行動パターンを抽出し，都市の特徴づけを行う手法を提案している．佐伯ら[7]は，訪日外国人に対する対面アンケートによる訪問先や感想の調査の代替手段として，位置情報付きツイートの分析による使用言語別や季節による訪問先の差異を分析している．我々は，位置情報付きツ

†1 職業能力開発総合大学校　Polytechnic University
†2 首都大学東京　Tokyo Metropolitan University
†3 産業技術総合研究所　National Institute of Advanced Industrial Science and Technology (AIST)
†4 大分工業高等専門学校　National Institute of Technology, Oita College

イートを分析することで，観光に関する各地域の特徴的な情報を自動抽出する手法を検討する．そして，手法を適用することで，旅行者が旅行時に収集する各地域の特徴であるご当地グルメなどを低コストに取得する手法を提案する．

3. 提案手法

観光情報を収集する際，旅行先地域の特徴となるグルメやイベントを知ることができれば，旅行者が地域の特徴を Web 検索により取得するコストを削減でき，より有用な観光情報収集が可能となる．また，情報弱者の情報への不到達を防ぎ，対象地域の観光情報として有用な情報を提示することができる．本稿では，一般的な料理名(例：うどん，そば)やイベント名(例：祭り，踊り)のみで Twitter を分析することによって，旅行先地域の特徴抽出を行う．

図 1 にシステム概要を示す．システムは，①データ収集，②前処理，③特徴抽出，④出力の処理で構成される．各処理について以降の各節で記述する．

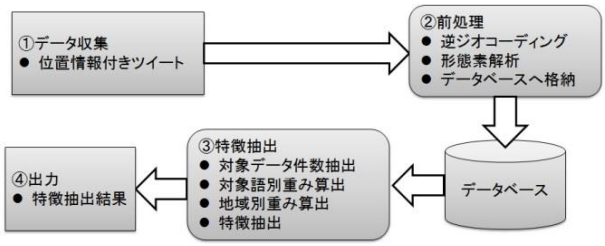

図 1 システム概要

3.1 データ収集

本節では，図 1 に示した①データ収集の手法について記述する．本稿での目的は，観光情報としての地域の特徴を抽出することである．そのため，位置情報が特定可能なツイートを利用した分析によって，地域別の特徴を抽出する必要がある．そのため，我々は，発信された地域を特定可能である Twitter 内の位置情報付きツイートに限定した分析を行うこととした．

収集対象は，Twitter から発信された位置情報付きツイートの中で，日本の領土を含む範囲である緯度経度が 10 進法表記で 120.0 ≦経度≦154.0 かつ 20.0≦緯度≦47.0 である位置情報付きツイートとした．位置情報付きツイートのデータの収集には，Twitter 社が提供する API の 1 つである Streaming API[8]を用いた．

次に，収集したデータ数について述べる．橋本ら[9]の研究によると，日本国内で発信されるツイートの中で位置情報が付いている割合は約 0.18%とツイート全体では非常に少ないデータ数である．しかし，収集した位置情報付きツイートは，表 1 に示す推移例のとおり，平日でも約 7 万件，土日には 10 万件を超える日もある．本研究において収集した位置情報付きツイートは，2015/2/17 から 2016/2/16 までの期間で約 2,500 万件である．また,期間中の 1 日当たりの収集件数は約 67,000 件であった．このデータセットを用いて次節以降で述べる処理により地域別の特徴抽出を行った．

表 1 位置情報付きツイートの推移例 (2015/5/9-6/3)

日付(曜日)	件数	日付(曜日)	件数
5/9(土)	117,253	5/22(金)	92,237
5/10(日)	128,654	5/23(土)	55,590
5/11(月)	91,795	5/24(日)	72,243
5/12(火)	87,354	5/25(月)	82,375
5/13(水)	67,016	5/26(火)	83,851
5/14(木)	88,994	5/27(水)	83,825
5/15(金)	89,210	5/28(木)	85,024
5/16(土)	116,600	5/29(金)	121,582
5/17(日)	126,705	5/30(土)	119,387
5/18(月)	89,342	5/31(日)	81,431
5/19(火)	83,695	6/1(月)	76,364
5/20(水)	87,927	6/2(火)	76,699
5/21(木)	86,164	6/3(水)	78,329

3.2 前処理

本節では，図 1 に示した②前処理について記述する．3.1 節の処理により収集したデータに対して，逆ジオコーディング・形態素解析・データベースへの格納の処理を行う．

逆ジオコーディングは，収集した個々のデータ(ツイート)の緯度経度情報から都道府県，(郡)市町村・特別区，町・字を特定した．これは，独立行政法人農業・食品産業技術総合研究機構の簡易逆ジオコーディングサービス[10]を用いた．例として，(緯度,経度)=(35.7384446,139.460910)を逆ジオコーディングすると，都道府県名：東京都，(郡)市町村・特別区：小平市，町・字：小川西町二丁目が得られる．本稿では，地域別の特徴抽出を行うための地域単位として，都道府県別と市区町村別を想定している．(八地方区分，都道府県内の地域区分は別途検討)

形態素解析は，収集した個々のデータ(ツイート)の本文を形態素解析器「MeCab」[11]を用いて，分かち書きを行う．例として，「桜がきれいです。」は，「桜/名詞，が/助詞，きれい/名詞，です/助動詞，。/記号」と分割される．なお，本稿の提案手法では， Wikipedia の見出し語ファイル[12]を利用し，Wikipedia に登録されている見出し語を予め MeCab のユーザ辞書へ追加している．

データベースへの格納は，データ収集・逆ジオコーディング・形態素解析の処理を行った結果から特徴抽出に必要となるデータをデータベースに格納する．本研究において利用したデータは，ツイート ID・ツイート本文・形態素解析結果・緯度・経度である．

3.3 特徴抽出

本節では，図 1 に示した③特徴抽出について記述する．特徴抽出は，収集した位置情報付きツイートから地域別の特徴を自動抽出する手法である．

特徴抽出を行うためには，分析対象とする範囲が必要である．そのため，前節で述べた地域別(都道府県別，地区町村名別)を基準に行うこととした．

また，特徴抽出は，前処理により生成したデータに対し，分析対象となる対象語の出現頻度を基準に分析を行う．そのため，分析対象とする対象語が必要となる．この対象語は，観光情報を取得したいユーザが決めることを想定している．本稿では，提案手法の有用性を確認するために予め指定することとした．

対象語の指定には，基準語として「うどん」や「そば」，「祭」などの一般的な食やイベントに関する語を決める．次に，基準語で形態素解析結果を検索し，各基準語が含まれる語を特徴抽出の対象語として選択する．選択された各基準語が含まれる対象語一覧を対象語群と定義し，各対象語群別に分析を行うこととした．

以降に，対象語群ごとに地域別の特徴抽出を行う手順を示す．まず，前処理により分割された形態素に対象語群が含まれるデータについて，対象語別および地域別の件数を抽出する．対象データ件数は，(1)・(2)・(3)式で示す．(3)式における$c_{w_X p_Y}$は，対象語「w_X」の地域「p_Y」での出現回数を表す．

対象語群$W = \{w_1, w_2, w_3, \dots, w_X\}$　・・・(1)

地域群$P = \{p_1, p_2, p_3, \dots, p_Y\}$　・・・(2)

対象データ件数$C(X \times Y) = \begin{bmatrix} c_{w_1 p_1} & c_{w_1 p_2} & \cdots & c_{w_1 p_Y} \\ c_{w_2 p_1} & c_{w_2 p_2} & \cdots & c_{w_2 p_Y} \\ \vdots & \vdots & \ddots & \vdots \\ c_{w_X p_1} & c_{w_X p_2} & \cdots & c_{w_X p_Y} \end{bmatrix}$

・・・(3)

次に，対象データ件数について対象語別の重みを(4)式により算出する．同様に，地域別の重みを(5)式により算出する．(4)・(5)式を用いて(6)式により対象語別・地域別のスコアを求め，特徴抽出に利用した．

$$D_{w_i p_j} = \frac{c_{w_i p_j}}{max\left(c_{w_1 p_j}, c_{w_2 p_j}, \cdots c_{w_X p_j}\right)}$$ ・・・(4)

$$E_{w_i p_j} = \frac{c_{w_i p_j}}{max\left(c_{w_i p_1}, c_{w_i p_2}, \cdots c_{w_i p_Y}\right)}$$ ・・・(5)

$$Score_{w_i p_j} = D_{w_i p_j} \times E_{w_i p_j}$$ ・・・(6)

以降に示す実験では，(6)式により算出した *Score* を基準に値が高い対象語を地域の特徴と捉え抽出を行った．

3.4 出力

本節では，図 1 に示した④出力について記述する．出力は，前節までの処理により特徴抽出を行った結果を利用した可視化を想定している．本稿では，横軸に *Score*，縦軸に対象語を取ったグラフと地図を利用した *Score* 分布の可視化例を示した．旅行時に有用となる可視化手法については，今後の課題とする．

4. 実験方法と実験結果

本章では，3 章で述べた提案手法を用いた特徴抽出の実験について記述する．4.1 節に実験に利用したデータセットを示し，4.2 節に食に関する実験，4.3 節にイベントに関する実験を記述する．

4.1 データセット

本実験で使用したデータセットは，3.1 節のデータ収集で述べた Streaming API を用いて収集した，2015/2/17 から 2016/2/6 までの期間の日本国内の緯度経度情報を含む位置情報付きツイートである約 2,500 万件とした．このデータセットを用いて特徴抽出を行った．

本稿では，観光情報として旅行時に必要と想定される食とイベントを特徴抽出の実験対象とした．4.1 節に食，4.2 節にイベントの実験について記述する．

4.2 食に関する実験

本節では，4.1 節のデータセットを利用し，食に関連した特徴抽出を行った実験について記述する．ここで，我々は，位置情報付きツイートを分析することで，観光情報として有用なご当地グルメなど地域の特徴抽出を目的としている．そのため，観光振興として各地域で盛んに取り組まれているご当地グルメや B 級グルメで使われる語である「うどん」・「そば」・「ラーメン」を例に実験を行った．本研究では最終的に基準語をシステム利用者が入力可能とすることを目的としているが，本稿では予め定義し実験を行っている．

表 2 に対象語群とデータ数を示す．表 2 に示すとおり，基準語「うどん」・「そば」・「ラーメン」の平仮名表記・片仮名表記・漢字表記が含まれる語を対象語群とした．各対象語群に含まれる対象語群(例)について，データ件数が多い上位 5 件をデータ数と共に示す．また，対象語群に含まれる対象語数と各対象語群の全データ数も表 2 に示した．

なお，本稿では，例として，「ラーメン」を含む語である「富山ブラックラーメン」は抽出対象であるが，「ラーメン」を含まない「富山ブラック」は抽出対象としていない．対象語群として，「富山ブラックラーメン」と「富山ブラック」は同義語とする処理が必要となるが，本稿ではこの処理については言及しない．

表 2 対象語群とデータ数

基準語	対象語群(例)	データ数[件]	対象語数[語]（全データ数[件]）
うどん ウドン 饂飩	うどん	23,966	134 (45,968)
	はなまるうどん	2,820	
	カレーうどん	2,509	
	讃岐うどん	2,047	
	うどん屋	1,389	
そば ソバ 蕎麦	そば	34,739	376 (97,678)
	蕎麦	11,588	
	中華そば	10,924	
	油そば	8,406	
	焼きそば	3,704	
ラーメン らーめん 拉麺	ラーメン	98,700	212 (203,945)
	らーめん	39,030	
	ラーメン二郎	13,657	
	ラーメン屋	7,202	
	家系ラーメン	6,030	

表 2 に示すとおり，各対象語群でデータ数が最も多くなる対象語は「うどん」・「そば」・「ラーメン」となり，この対象語のみでは地域の特徴を捉えることは困難である．同様に，地域別の特徴をデータ数で求めた場合も，データ数の多い都市部が優位となる．

そこで，我々は，各対象語群の対象語について，前節の(6)式を用いた *Score* を算出し，地域の特徴を捉えることとした．以降に各対象語群での実験結果を示す．

4.2.1 対象語群「うどん」の実験結果

図 2 に，富山県における基準語「うどん」(片仮名・漢字表記も含む)の分析結果を示す．ここでは，算出した *Score* の値が高い上位 10 件を示した．なお，横軸は，*Score* を最大値で規格化し，ログスケールで表す．

図 2 において最上位に抽出された対象語は，富山県のご当地うどんである「氷見うどん」となった．また，第 3 位の「カレーうどん」も，富山県内では有名な店に関するツイートによって上位に抽出される結果となった．

次に，提案手法による *Score* と比較するため，表 2 の対象語別のデータ数を用いた特徴抽出について，上位 10 件を図 3 に示す．図 3 のデータ数を用いた場合，最上位は対象語「うどん」となり，第 2 位に「カレーうどん」，第 3 位に「氷見うどん」となった．

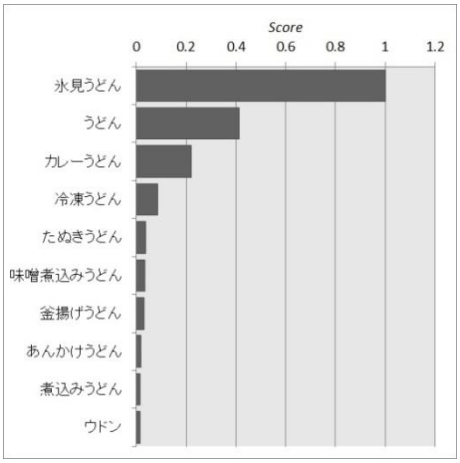

図 2 富山県での基準語「うどん」の分析結果

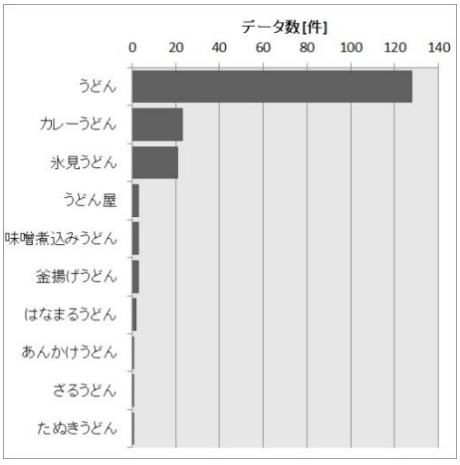

図 3 データ数による富山県での基準語「うどん」の分析結果

図 2 と図 3 の実験結果から，「うどん」は一般的に用いられる語であるためデータ数が多く，データ数のみで地域の特徴を捉えることは困難である．また，「カレーうどん」と「氷見うどん」を比較した場合も，より地域の特徴を表す「氷見うどん」は，「カレーうどん」のデータ数を下回り，地域の特徴として埋もれてしまう可能性がある．しかし，提案手法を利用した図 2 では，地域の特徴抽出に成功した．よって，本稿の提案手法による *Score* を利用することで，より地域の特徴を表す対象語を上位に抽出できる．

次に，図 2 で上位に示された対象語について，都道府県別の分布を示す．この処理によって，旅行者は抽出された基準語が旅行先地域の特徴を表す語であるかを判断し，ご当地グルメを確認する一助となる．本稿では，Microsoft Power BI Desktop[13]を利用した可視化により確認を行った．ここで，図 4 に，図 2 において最上位に抽出された対象語「氷見うどん」の都道府県別分布を示す．図は，*Score* が高い都道府県ほど濃く表される．「氷見うどん」は，47 都道府県中の 5 都県のみで抽出され，図 4 において最も濃く示されており，富山県特有のご当地グルメと判断できる．

また，図 5 に対象語「カレーうどん」の都道府県別分布を示す．「カレーうどん」については，全都道府県で抽出された．富山県での「カレーうどん」の *Score* は 47 都道府県中で 14 位となり，富山県特有の特徴とは言えない．しかし，図 2 の富山県内での *Score* は第 3 位となり，富山県内でうどんを選択する際に，「カレーうどん」も候補の 1 つと判断できる可能性を示した．なお，「カレーうどん」は，茨城県が最も高い *Score* を示しており，茨城県内に全国的に有名な「カレーうどん」の店の存在を示唆する．

図 4 対象語「氷見うどん」の都道府県別分布

図 5 対象語「カレーうどん」の都道府県別分布

4.2.2 対象語群「そば」の実験結果

図 6 に，富山県における基準語「そば」の分析結果について，*Score* の値の高い上位 10 件を示した．最上位の対象語は，富山県のご当地そばである「利賀そば」となった．また，第 3 位には「山菜そば」が抽出された．

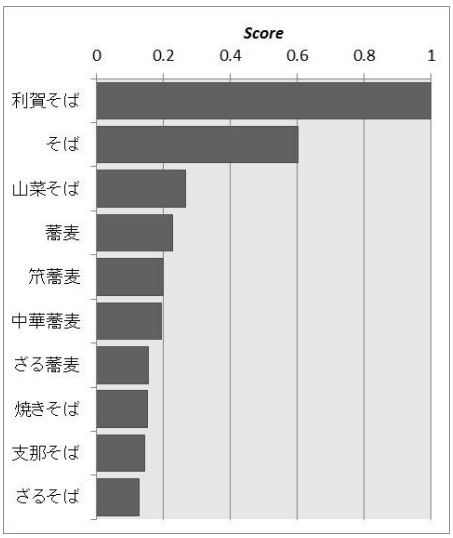

図 6 富山県での基準語「そば」の分析結果

次に，図 6 において上位に示された語について，都道府県別の分布を示す．図 7 に，図 6 で最上位に抽出された対象語「利賀そば」の都道府県別 *Score* を示す．「利賀そば」は，47 都道府県で富山県のみ抽出された．そのため，富山県の特徴であると判断できる．

図 8 に，対象語「山菜そば」の都道府県別分布を示す．「山菜そば」は，中日本を中心に全国に広く分布し，長野県が最も高い *Score* でとなった．富山県も 47 都道府県内で第 2 位の *Score* となっているため，都道府県別の *Score* からも「山菜そば」は，富山県特有のご当地グルメである可能性を示している．

図 7 対象語「利賀そば」の都道府県別分布

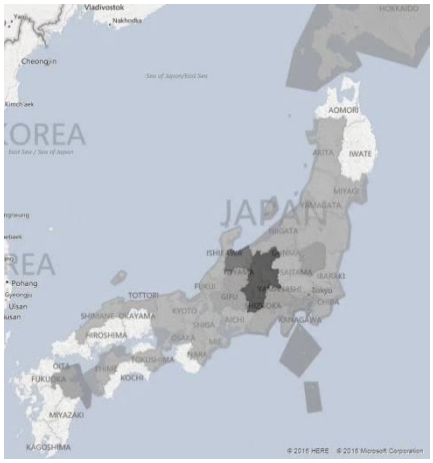

図 8 対象語「山菜そば」の都道府県別分布

4.2.3 対象語群「ラーメン」の実験結果

図 9 に，富山県における基準語「ラーメン」の分析結果について， *Score* の値が高い上位 10 件を示した．最上位の対象語は，北陸地方を中心に展開しているラーメンチェーン店の「8 番らーめん」となった．第 3 位は，片仮名表記の「8 番ラーメン」となった．また，第 5 位・第 6 位は富山県内のラーメン店名，第 7 位・第 8 位はご当地ラーメン，第 9 位は全国展開をしているラーメンチェーン店であった．

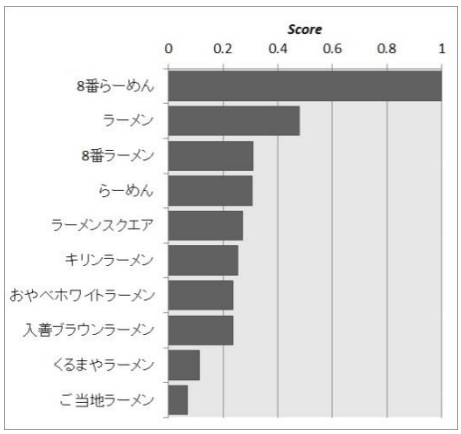

図 9 富山県での基準語「ラーメン」の分析結果

次に，各対象語の都道府県別分布を示す．図 10 に最上位に抽出された対象語「8 番らーめん」，図 11 に第 2 位の「8 番ラーメン」を示す．図 10 と図 11 において，最も *Score* が高い地域は共に石川県であり，チェーン店を展開する府県を中心に分布を確認できる．

ここで，対象語「8 番らーめん」と「8 番ラーメン」を (7)式に示す *Cosine* 類似度を利用し比較した．(7)式でのVは，47 都道府県を意味する．X_aを「8 番らーめん」，X_bを「8 番ラーメン」として，各都道府県の *Score* を利用し類似度を求めた結果，類似度は 0.92 となった．このため，対象語である「8 番らーめん」と「8 番ラーメン」は，表記揺れがあるものの同義語であると考えられる．本稿では，地域の特徴抽出に焦点をあてているため同義語の判定は行っていないが，本研究の次の段階として，推定される地域の特

徴抽出の精度向上には，同義語判定も必要になると考えられる．

図 10 対象語「8 番らーめん」の都道府県別分布

図 11 対象語「8 番ラーメン」の都道府県別分布

$$cos(\overrightarrow{Score_{X_a}}, \overrightarrow{Score_{X_b}}) = \frac{\sum_{i=1}^{|V|} Score_{X_a p_i} Score_{X_b p_i}}{\sqrt{\sum_{i=1}^{|V|} Score_{X_a}{}^2} \cdot \sqrt{\sum_{i=1}^{|V|} Score_{X_b}{}^2}} \quad \cdots (7)$$

図 12 に，図 9 の第 5 位の対象語「ラーメンスクエア」の都道府県別分布を示す．東京都の「ラーメンスクエア」は，立川市にあるテーマパーク型のラーメンのフードコートであり，最も *Score* が高い結果となった．富山県の「ラーメンスクエア」は，ラーメン店の店名であり，第 1 位・第 3 位のチェーン店名である「8 番らーめん」と比較すると *Score* は低いが，「ラーメン」を含む店名の中では富山県内において特徴的なラーメン店である．

同様に，図 13 に対象語「キリンラーメン」の都道府県別分布を示す．「キリンラーメン」は，他県にも抽出されているが，富山県が最も *Score* が高い結果となった．よって，対象語「キリンラーメン」として最も特徴的な都道府県は富山県であると考えられ，富山県特有のラーメン店の候補となる可能性を示した．

図 12 対象語「ラーメンスクエア」の都道府県別分布

図 13 対象語「キリンラーメン」の都道府県別分布

図 14 に，対象語「おやべホワイトラーメン」の都道府県別分布，図 15 に対象語「入善ブラウンラーメン」の都道府県分布を示す．ここで示した 2 つのラーメンは，町おこしのために，富山県小矢部市と富山県下新川郡入善町に

図 14 対象語「おやべホワイトラーメン」の都道府県別分布

おいて，近年作られたご当地ラーメンである．富山県以外には分布がなく，ご当地グルメとして考案されたラーメンを地域の特徴として正しく抽出し，富山県のご当地グルメを捉えている．

図 15 対象語「入善ブラウンラーメン」の都道府県別分布

4.3 イベントに関する実験

本節では，4.1 節のデータセットを利用したイベントの抽出について記述する．前節では，食に関する実験について記述し，ご当地グルメの抽出を行った．ここでは，提案手法の適用により，旅行対象とした地域の特徴となるイベントの抽出実験を行った．

本実験における対象語は，基準語「まつり」(祭を含む)が含まれる語とした．4.2 節の表 2 に示した食に関する実験と同様に，イベントに含まれる可能性のある基準語を予め人手で定義し，基準語が含まれる語を対象語とした．

表 3 に，イベントに関する実験で利用した対象語群とデータ例を示す．各対象語群に含まれる対象語例について，データ件数が多い上位 5 件をデータ数と共に示す．また，対象語群に含まれる対象語数と各対象語群の全データ数も併せて示した．

表 3 に示した対象語群の例のとおり，データ数が多い語は，「祭り」をはじめ「文化祭・体育祭」など，地域によらず日常的に用いられる機会が多い語である．そのため，食の実験と同様に，対象語のデータ数だけでは地域の特徴を抽出することができない．そこで，我々は 3 章で述べた提案手法を用いてイベントに関する特徴抽出を試みた．以降で実験結果を示す．

表 3 対象語群とデータ数

基準語	対象語群(例)	データ数[件]	対象語数[語] (全データ数[件])
まつり 祭	祭り	15,229	844 (80,104)
	まつり	5,111	
	お祭り	4,303	
	文化祭	4,149	
	体育祭	2,672	

4.3.1 イベントに関する対象語での実験結果

図 16 に，富山県における表 3 に示した基準語「まつり」の分析結果を示す．ここでは，算出した *Score* の値が高い上位 10 件を示した．なお，横軸は，*Score* を最大値で規格化し，ログスケールで表す．

図 16 において，上位に抽出された対象語は，富山県内で有名な祭となった．よって，提案手法の適用により「まつり」も地域の特徴抽出が可能である．

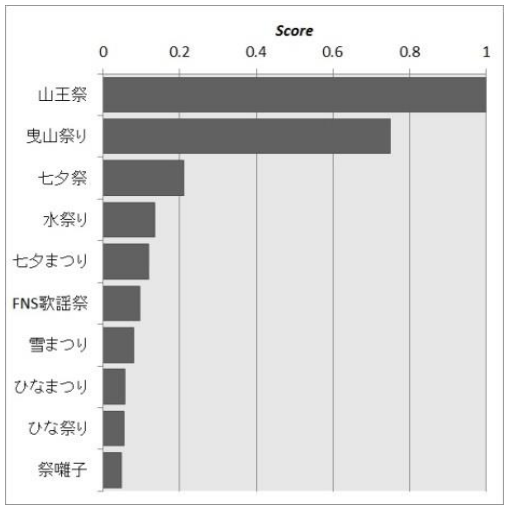

図 16 富山県での基準語「まつり」の分析結果

次に，図 16 の上位 3 件について，各対象語の都道府県別分布を記述する．

図 17 に対象語「山王祭」，図 18 に対象語「曳山祭り」についての分布を示す．図 17 の「山王祭」と図 18 の「曳山祭り」については，47 都道府県中で最も *Score* が高い結果となった．この 2 つの祭は，他地域でも同様の名称の祭は存在するが，提案手法により富山県に最も特徴があると抽出した．

図 17 対象語「山王祭」の都道府県別分布

図 19 に対象語「七夕祭」についての分布を示す．図 19 の「七夕祭」は，全国的に有名な宮城県仙台市の「(仙台)

七夕祭」が最も *Score* が高い結果となった．次いで，富山県高岡市の「(戸出)七夕祭」が 47 都道府県中で第 2 位，第 3 位に神奈川県平塚市の「(湘南ひらつか)七夕祭」が抽出された．この結果から，対象語「まつり」について，都道府県単位での地域の特徴を捉えている．

図 18 対象語「曳山祭り」の都道府県別分布

図 19 対象語「七夕祭」の都道府県別分布

5. まとめ

本稿では，Twitter の位置情報付きツイートを用いて各地域における観光情報に関連するご当地グルメやイベントを取得するために，位置情報付きツイートに含まれる対象語を地域別に分析する手法を提案した．本稿の実験では，食に関して「うどん」・「そば」・「ラーメン」，イベントに関して「まつり」について述べた．富山県での分析結果から，富山県内における地域の特徴を捉えた．また，都道府県別分布の分析結果から，富山県特有であるか他地域にも存在する食・イベントであるかも捉えた．提案手法を用いた地域特徴抽出を利用することで，地域のご当地グルメやイベントを抽出し，観光情報を提供する上での一助となる可能性を示した．

今後は， SNS の即時性・網羅性をさらに活用するため，時系列情報を考慮し，季節ごとに該当地域の流行を取得する手法を検討する予定である．また，旅行先地域に関する知識がない場合であっても，旅行先の観光情報を取得可能であるかをさらなる評価を行う計画である．

参考文献

[1] Twitter, https://twitter.com/ (2014).
[2] 首相官邸公式アカウント @kantei, https://twitter.com/kantei (2011).
[3] 土屋 圭, 豊田 正史, 喜連川 優, “マイクロブログを用いた鉄道の運行トラブル状況抽出に関する一検討”, 信学技報，Vol.113, No.150, DE2013-30, pp.175-180 (2013).
[4] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo, “Earthquake shakes Twitter users: real-time event detection by social sensors”, WWW2010, pp.851-860 (2010).
[5] Ryong Lee, Kazutoshi Sumiya, “Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection”, Proc. 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, LBSN ’10, pp.1-10 (2010).
[6] 李 龍, 若宮 翔子, 角谷 和俊, “Tweet 分析による群集行動を用いた地域特徴抽出”, 情報処理学会 データベース，Vol.5, No.2, pp.36-52 (2012).
[7] 佐伯 圭介, 遠藤 雅樹, 廣田 雅春, 倉田 陽平, 石川 博, “Twitter データを利用した訪日外国人の訪問先の言語別分析”, 観光情報学会論文誌 観光と情報, 第 11 巻, 第 1 号, pp.45-55 (2015).
[8] Twitter Developer 公式サイト, https://dev.twitter.com/ (2014).
[9] 橋本 康弘, 岡 瑞起, “都市におけるジオタグ付きツイートの統計”, 人工知能学会誌, 27 巻, 4 号, pp.424-431 (2012).
[10] 独立行政法人農業・食品産業技術総合研究機構:簡易逆ジオコーディングサービス，http://www.finds.jp/wsdocs/rgeocode/index.html.ja (2014).
[11] MeCab:Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html (2012).
[12] Wikipedia: データベースダウンロード，https://ja.wikipedia.org/wiki/Wikipedia:データベースダウンロード (2014).
[13] Microsoft Power BI Desktop, https://powerbi.microsoft.com/ja-jp/desktop/ (2016).