ツイッターマイニングによる ソーシャルイベントの影響度の推定と実空間との関連性の調査

藤田 哲也　　杉山 歩　　鈴木 健之
山口 和宏　　Dam,Hieu Chi　　Ho,Tu Bao
北陸先端科学技術大学院大学 知識科学研究科
{ s1050045,a-sugiya,takeyuki.s1050027
Kazuhiro_YAMAGUCHI,dam,bao}@jaist.ac.jp

1　序論

近年のMicroblogやSNSの急速な広まりは，研究・ビジネス分野の両面で幅広い展開と展開が期待されている．中でも2004年にアメリカの青年Mark Elliot Zuckerbergらが開発したFacebookは2011年11月時点において全世界で11億人もの利用者が存在し，近日上場される株価は1000億ドル(7.7兆円)を超えるものと予想されている．またFacebook同様アメリカで2006年に誕生したTwitterは最大140字という文字制限が与える気軽さが人気を博し，またスマートフォンなどのモバイル機器との親和性が高さから日本での利用者は1400万人と国内最大規模のSNSの一つとなった[1]．現在ではTwitterからは全世界で秒間3000回以上もの発信があり，メディア媒体のありかたを揺るがすほどの規模となっている．

Twitterの特徴は文字数制限以外にその即時性とリツイートと呼ばれる発言を拡散するシステムにも特徴がある．これらのシステムにより一人の発言は瞬く間に世界中のユーザーに広まるようになった．昨年度の東日本大震災時には多くの情報をユーザー間で共有することが可能となり，マスメディア，電話と同様，社会インフラの一つとして認知されつつある．一方，そのリアルタイムでの膨大なツイート数は逆にそのツイート数から社会での動きやイベントを特定することも可能である．昨年のサッカー女子日本代表によるワールドカップ優勝時には秒間7916回ものツイートがあり，多くのユーザーがその喜びを共有していたことが特定され，先日のアメリカンフットボール，スーパーボウル開催時にはアメリカで秒間4064回とこれも多くのアメリカ人が一つの話題を共有していたことがわかる．さらに，リツイート等のシステムによりユーザー間でのイベント告知の共有はサッカーなど従来までの既知の国民的関心事以外のイベントにも大きな影響力を持たせることもある．先日のTVで放送された映画は20年以上前に公開されたもににも関わらずユーザー間の告知効果からピーク時には秒間2万5000回ものツイートが起こり，SNSやネットワーク世代での関心事は新たな展開を持っているものと考えられる．

この映画やスポーツ中継の事例が示す様にネットワーク空間での関心事と現状と既存のメディアの関心事は一部では共有可能であるが，一部ではまったく違った様相を呈するイベントもある．SNSの観点からイベントの可能性を掘り起こす事はビジネス上興味深く，イベントの参加人数の少なさから新聞の記事では些細な扱いとなったり，記事にすらならないイベントでもSNSでは大きなイベントであったり，極ローカルなイベントと思われていたイベントがSNS上では全国的に関心を示す人々が多いイベントである事などを特定することは今後のイベントの重要度を示す尺度の一つとなるものと考えられる．またSNS上のコミュニティーを理解する上でも興味深い．

そこで本研究では，規模，対象ユーザー，場所などの条件の異なるいくつかのイベントを対象とし，ユーザーの位置情報から定義されるつぶやき集合の差からイベントの影響度と実空間上での相関性について考察をおこなう．これまでの先行研究でもGeotagや位置情報を利用した解析ではイベントの検出・位置推定などで成果を上げており，そこにイベントの性格やユーザー属性の違いを考慮に入れた影響度の測定を行い，SNS上でのイベント毎の性格の違いについて報告する[2, 3]．

2 手法

本研究では twitter のツイートデータを取得し，そのユーザープロファイルからテキスト処理により発信者の位置を特定システムを構築する．本研究で利用する位置情報は発信者のリアルタイムでの所在地では無く，発信者が普段生活している所在地を基本とし，またその位置情報は都道府県別に集計をおこなう．以下にデータの取得並びに解析手法について説明する．

2.1 データ取得方法

データ取得には PTT(Python Twitter Toolset) ライブラリを用いて Search API，並びに REST API(show/lookup メソッド) から時間毎にデータを集計し，ユーザープロファイルを取得する [4]．Search API では特定のキーワードをクエリとして与えることでツイート情報を取得できる．

ここで保存するツイートデータはツイート ID，投稿時間，screen_name，本文である．screen_name はユーザーを一意に識別するユーザー ID であり，現在最大 15 文字の英数字からなる．Search API を利用する場合，ユーザーの名前以外のユーザープロファイルは取得できないため REST API を用いて，screen_name からプロファイル情報を取得する．今回はこの中に含まれる位置情報である location データを用いる．location データはユーザーが任意に設定できる最大 30 文字からなる文字列である．

2.2 位置情報決定方法

ユーザープロファイルに記載されている位置情報は記述方法が様々でかつ実在しない都市名も多く見受けられる．本研究ではこれらの紛らわしい地域名，都市名を排除するために以下の手順をおこなった．まず，データの特徴を把握するためにランダムに 1000 件出した位置情報に手作業でクラス分けをおこなった．その際，地域粒度として都道府県 (以降，県と表記) 単位までを情報抽出の対象とし，個別のクラスとした．

データのクラス分けをおこなった結果，県名・県庁所在地などが頻出し，その際には「市区町村」など地域接尾辞が省かれている．英語での県名表記，それ以外の地域名では地域接尾辞を省いていない等の特徴が見られたことから，辞書を用いた文字列検出手法とパターンマッチングによる地域検出手法，二つの位置同定手法を用意して位置検出をおこなった．

文字列検出手法では県名など特定ワードの辞書を作成し，その特定の文字列に該当すれば所属する県と関連づける．関連づけられた地名が複数合った場合はその他のクラスに振り分けることとした．辞書の作成時，英語表記を記述する際には，ローマ字表記，表記揺れを対応した．

パターンマッチング手法では，位置情報を○○市などの地域名＋特定キーワードで抽出する．その際，記号文字，平仮名，片仮名を区切り文字として扱い，元の位置情報文字列を分割して個別の文字列に対してパターンマッチをおこなう．また，抽出された地域名は郵便事業株式会社が公開している郵便番号と住所を関連づけた住所録を用いて，市町村名から県の検索をおこなう．今回は単一の県が検出された場合にのみ位置情報を確定させている．

文字列検出手法で検出ができた場合，パターンマッチング手法は適用しない．また，両者の手法で該当しなかった場合はその他のクラスに振り分ける．

2.3 イベントの影響度測定方法

本研究は 2011 年 12 月に行われた参加者が 10 万人以上いる規模の大きなイベントの中から 3 つのイベントとして，クリスマス，Sony PSVITA 発売 (以降，VITA)，コミックマーケット 81(以降，コミケ) を注目し，それぞれ関連するキーワードを含む tweet データを収集し，ツイートの投稿者をイベント毎のユーザー集合として扱った．

その後，投稿者のユーザープロファイルを取得し，ツイート投稿者の本拠地の情報を位置同定手法を用いてユーザー毎に付加した．

本研究の解析に利用したイベントを表 1 に示す．

表 1: イベントの種類とイベント属性

イベント名	イベントの種別	規模	開催地
クリスマス	全世代	国民的	全国
PSVITA 発売	若年層	10 万人台/日	全国
コミケ 81	若年層	10 万人台/日	東京

表 2: クラス分類と分類結果の例

location	クラス	結果	正誤
茨城	茨城県	茨城県	正
神奈川県茅ヶ崎市	神奈川県	神奈川県	正
Osaka, Japan	大阪府	大阪府	正
だ埼玉	埼玉県	埼玉県	正
流山市	千葉県	千葉県	正
東京，たまに茨城だっぺよー!	その他	その他	正
京都市在住，大阪市在勤	京都府	その他	誤
彩の国.	埼玉県	その他	誤
23 区	東京都	その他	誤
くにたち市	東京都	その他	誤

表 3: イベント毎の tweet ユーザー数と特定された本拠地数および集計日

イベント名	ユーザー総数	本拠地付数	集計日
クリスマス	1034938	642890	2011/12/23,24,25
PSVITA	87429	34454	2011/12/16,17,18
コミケ 81	240938	77184	2011/12/29,30,31

3 結果と考察

3.1 テストデータに対する位置情報の決定精度

2.1 でクラス分けしたデータを用いて開発した手法の精度を計測したところ，分類正解率は 0.919 であった．正誤判定の一部を表 2 に示す．

3.2 Twitter からの位置情報の決定精度

2.2 で述べた手法により収集された tweet から発言者の本拠地を決定した．まず，イベント毎のキーワード検索を実行し，本拠地を特定したユーザー総数を表 3 に示す．本拠地を特定したユーザー数のうち全国全年齢に共通のイベントである「クリスマス」のユーザー数を比較すると，各都道府県の人口構成と近い傾向を示しており，本データがイベントから精度良く本拠地付きユーザーを割り出している事がわかる．

次に，3 つのイベントにおいて最大ユーザー数を誇る東京を基準に各イベントのユーザー数を正規化したものを図 1 に示す．この時,「VITA」キーワードの比率が「クリスマス」キーワードを上回った都道府県は関東に属する千葉県，神奈川県，埼玉県，群馬県，茨城県および愛知県であった．同様に「コミケ」キーワードが「クリスマス」キーワードを上回った都道府県は関東に属する都道府県のみであった．若年層に人気の高い「VITA」や「コミケ」に関東周辺部では強い関心を示す集団が多いことがわかる．

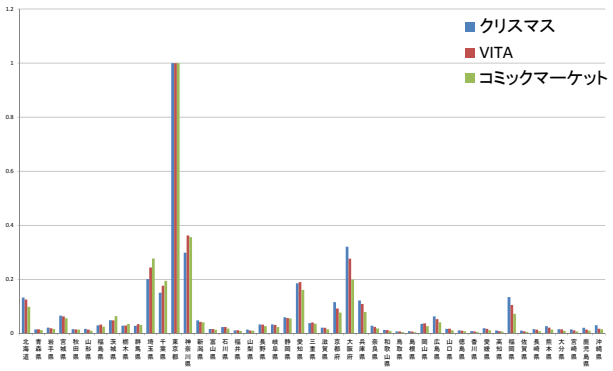

図 1: 東京を基準とした，各イベントの影響比率の変化

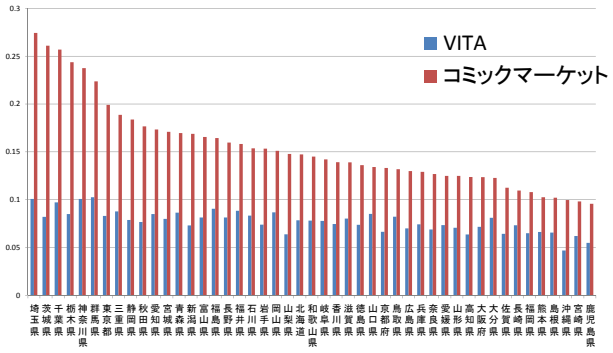

図 2: 県毎の人口比率に従ったユーザー数の影響比率の変化 (各県をクリスマスの値で正規化，コミケの値でソート)

3.3 イベントの種類と関心を持つ人の場所の相関性

ここでは，イベントの地域性とその範囲についてより詳細な検証をおこなった．前節で「クリスマス」,「VITA」,「コミケ」のイベント規模とツイート総数，本拠地付きユーザー数には相関性があることが示された．次にイベントごとの県別ユーザー数を比較した．ユーザー数を各県毎の人口比で標準化し，現実の世界での影響力に変換した．ここで，全国かつ全世代に関係するイベントである「クリスマス」を基準とし，「VITA」及び「コミケ」の影響力を示した．より若年層の多い 8 大都市を擁する都道府県において「クリスマス」よりも「VITA」,「コミケ」に対する関心が高いことがわかる．

また東京で開催された「コミケ」についてはこれら 8 大都市の場合であっても関西圏以西では関心は低く，東京都からの離れるにつれ関心が低いことがわかった．一方，全国的なイベントである「VITA」では距離の影響が無いことがわかった．

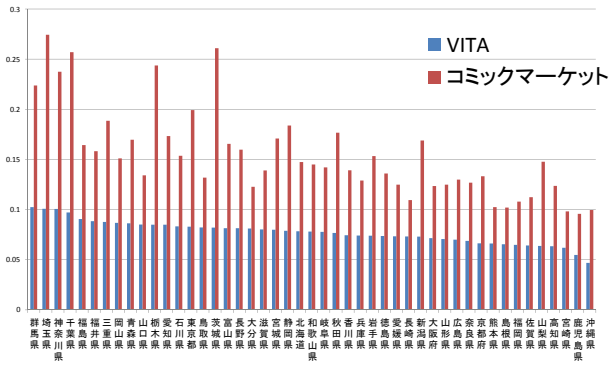

図 3: 県毎の人口比率に従ったユーザー数の影響比率の変化 (各県をクリスマスの値で正規化，VITA の値でソート)

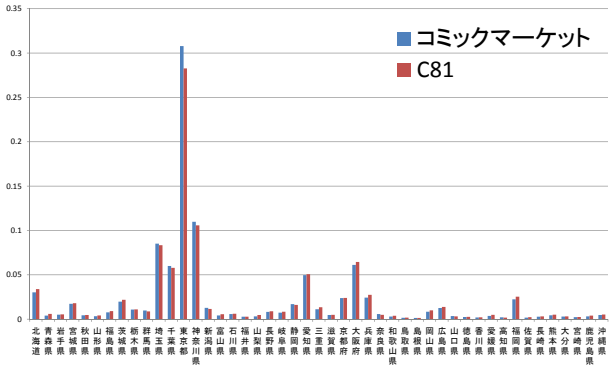

図 4: キーワードの違いによるユーザー層の差

さらに，さらに使用した検索キーワードの違いを調べるため，一般によく使われる「コミックマーケット」とコアなユーザーに使用される「C81」をつぶやいた発言者集合の比較をおこなった (図 4)．図 4 から東京都では他の都市と比べ,「C81」のキーワードを利用するユーザーの割合が小さく，東京ではコアなファン以外の層も幅広く関心を集めているといえる．以上の結果から東京開催のイベントでは距離に比例し，関心を持つ人口は減少しており，この減少率は関西以西で顕著であることと東京のユーザーは幅広い層が興味を持っている傾向にあることがわかった．

4 まとめ

本研究では twetter のユーザープロファイル情報をテキスト処理することでユーザーの本拠地を特定し，twitte の検索キーワードからイベントの影響度を判定するシステムを構築した．本システムによりイベントの対象と関心を持つ人の場所の相関性を評価し，全年齢が関心があるイベント (「クリスマス」) と比べ，若年層対象のイベント (「Sony PSVITA 発売」,「コミックマーケット 81」) では都市部を擁する都道府県において強い関心があることがわかった．また，開催地が一カ所である「コミックマーケット 81」では開催地からの距離に比例し関心を示す割合が弱くなることがわかった．さらに，利用キーワードの別を区別すると，短縮語の形 (「C81」) を利用するユーザーの割合は開催地 (東京都) 以外で多く，開催地での関心はコアなファン以外の層の厚さにあるものと考えられる．

今後，本システムを利用することで，規模や対象，開催地の異なる様々なイベントを評価し，イベントの種類別の傾向を特定していく予定である．また，現状のシステムでは位置情報の検出を限られたパターンに依存しているが，ユニークな表現や揺らぎのある表記に対応する予定である．また，ユーザープロファイル情報のみを利用しているが，今後は Geotag 等を用いてし，本拠地の特定精度と特定件数の向上を目指す．

参考文献

[1] twitter.com. Twitter.

[2] 藤坂達也, 李龍, 角谷和俊. 実空間マイクロブログ分析による地域イベントの影響範囲推定. In *DEIM Forum 2010 D7-4*, 2010.

[3] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. *Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors*, pp. 851–860. ACM, 2010.

[4] twitter. Twitter developers. `https://dev.twitter.com/`.