2E1-2in

ソーシャルメディアからの予告型の地域イベント及び参加状態の抽出手法の提案

Event Notification Extraction and Participants Estimation From Social Media

榊 剛史[*1] 那須野 薫[*1] 柳原 正[*2] 古賀 光[*2] 加藤 芳隆[*2] 那和 一成[*2]
Takeshi Sakaki Kaoru Nasuno Tadashi Yanagihara Kou Koga Yoshitaka Kato Kazunari Nawa

松尾 豊[*1]
Yutaka Matsuo

[*1]東京大学 The University of Tokyo
[*2]トヨタ IT 開発センター Toyota Info Technology Center

We propose an information extraction scheme to extract and provide scheduled geo-location based events to realize a richer experience to the driver. We apply a pattern matching method to extract both the scheduled geo-location based events and their participants from text-based messages obtained from social networking services. Experiments show that the method is capable of extracting the events with 0.745 accuracy and the participants with 0.864 accuracy.

1. はじめに

近年ではソーシャルネットワークサービス (SNS) に代表されるソーシャルメディアを解析し，他サービスに活用する取り組みが増えつつある．活用方法の一例として，人が関心を持ちやすい地域イベントの抽出に関連する取り組みが挙げられる [Watanabe 11, Hong 12]．抽出された地域イベントは，ユーザの行き先の推薦に活用することができる．

但し，このような推薦タスクにおいては，現在発生中のイベントではなく，これから起こることが予告されているイベント，すなわち予告型イベントが推薦候補として有意義であると言える．なぜなら，イベントがユーザから一定距離の範囲内に無ければ参加することが困難であり，仮に参加可能であっても，到達時にはイベントが完了している可能性があるためである．このような予告型イベントには花火大会や祭り，コンサート等，娯楽性の高いものが含まれる．

また，イベントそのものの情報を提供するだけでなく，現地の混み具合や参加状況などの状況も併せて提供することが望ましい．現地の状況は，そのイベントの参加者に問い合わせることで得ることができるが，このためにはイベントの参加者の抽出に加え，そのユーザの参加状態を特定する必要がある．

このような予告型地域イベントに関する情報は，従来雑誌やウェブページなどの静的なメディアとラジオやテレビなどの動的なメディアの両方で提供されてきた．前者は網羅的に情報を扱えるがリアルタイムに現地の状況を伝えることが困難である．後者はリアルタイムな情報が扱えるが同時に複数のイベントを扱うことが困難である．それに対し，ソーシャルメディアには大規模なユーザから，イベントについてのリアルタイムな情報が提供されている．それらを抽出することで，イベントに関するリアルタイムな情報を提供可能になると考えられる．

以上を踏まえ，本稿では下記の様なソーシャルメディアからの予告型の地域イベントの収集と，そのイベントに参加しているユーザと参加状態の抽出を行う手法を提案する．

予告型地域イベントの収集 予告型地域イベント及びその詳細な情報について予め収集するための手法．詳細な情報とは，開催場所やイベント名等である．

連絡先: 榊 剛史，東京大学大学院工学系研究科，東京都文京区弥生 2-11-16 工学部 9 号館

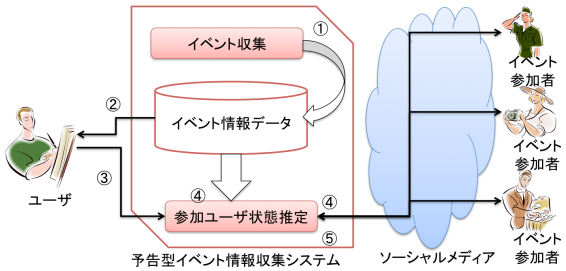

図 1: 予告型地域イベントの情報収集システム

ユーザのイベント参加状態推定 あるイベントについて参加しているユーザ集合を作成するために，各ユーザの参加状態を推定する手法．

最終的には，上記 2 つの手法を組み合わせて，図 1 のような「予告型地域イベント情報収集システム」を構築することを目指す．当該システムに関する詳細な説明は下記の通りである．

1. システム側で予め予告型地域イベントを収集しておく．
2. ユーザが利用する際には，現在地や目的地に合わせてそれらの周辺で行われているイベントを提示する．
3. ユーザが，現在状況を知りたいイベントを選択する．
4. 選択されたイベントへの参加ユーザを発見する．
5. 発見されたユーザの投稿から，イベント開催期間中の投稿を収集し表示する．

このような予告型地域イベント情報収集システムを構築することで，ユーザは様々なイベントについて現地の「生の声」を元に現在の状態を知ることができるため，より効率的な行き先の選択を実現することができると考えられる．

2. 関連研究

ソーシャルメディア上からイベントを抽出する研究は数多く行われている．これらの研究は 1. 特定のイベントを抽出する研究，2. 対象を絞らず，様々なイベントを抽出する研究の 2 つに分けることができる．

1. の研究においては，地震 [Sak 10] やインフルエンザの流行 [Aramaki 11]，スポーツの試合 [Chakrabarti 11] などを対象とした研究があげられる．これらにおいては，対象とするイベントの種類を絞ることでイベント発生検出や詳細な情報の抽出を可能にしている．そのため，様々な種類のイベントについて，汎用的に情報抽出を行うことは困難である．

2. の研究においては，話題になっているイベント [Lee 11] や，地域イベントを抽出する従来手法として [Watanabe 11] や [Hong 12] が挙げられるが，いずれも現在発生している地域イベントを対象としており，予告型のイベントを抽出できない．

本研究においては，予め対象とするイベントを自動的に大量に収集し，それぞれについて情報抽出を行う．そのため，様々なイベントを対象にすることができ，かつ各イベントについて詳細な情報を抽出することが可能になると考えられる．

3. 予告型地域イベント収集

予告型地域イベントを収集する手法について提案を行う．

予告型地域イベントとは，前もって開催される日時，場所，内容が告知されているイベントの事を指す．またイベント情報とは，そのイベントを一意に同定可能にする情報と定義する．本稿ではイベント情報は**開催日時**，**開催場所**，**イベント名**の 3 つにより構成されるものとする．普段，我々がイベントを同定する際も，この 3 つの要素を用いていると考えられる．そこで，本稿ではこれらを**イベント三要素**と呼ぶこととする．

本節では，イベント三要素を手掛かりに予告型イベントを収集する手法，また収集した予告イベントからイベント三要素を高精度に抽出する手法を提案する．

予告型地域イベントを収集する手法として 2 つの手法を適用する．一つはイベント情報サイトから収集する手法，もう一つはユーザ自身の発信情報を収集する手法である．

3.1 イベント情報サイトからの予告型地域イベント収集

旅行会社等により提供されているイベント情報ブサイトからイベント情報を収集する．イベント情報サイトとは，旅行会社や地方自治体等により提供されるイベント告知のためのウェブサイトを意味する．具体的には，「るるぶ.com」や[*1]，「地域情報サイト ZAQ」[*2]などがあげられる．

本稿では事例として，「るるぶ.com」からイベント情報を収集することを試みた．2012 年 1 月 1 日〜2012 年 12 月 31 日に，日本国内で開催されるイベントを全て収集した．結果，2268 件のイベントについてイベント情報を収集することができた．

また，これらのサイトにおいては，イベント情報を記述するフォーマットがサイトごとに決まっているため，各イベントから容易にイベント三要素を抽出することができる．

3.2 ユーザ発信情報からの予告型地域イベント収集

ソーシャルメディアからイベント情報を収集する手法を提案する．

図 2 のように，イベント主催者やイベントに興味を持つ個人がソーシャルメディア上にイベント情報を発信している．しかし，そのような発信方法ではユーザが自由にイベント情報を記述するため，単純な手法で収集することは困難である．収集したツイートからのイベント情報抽出も容易ではない．

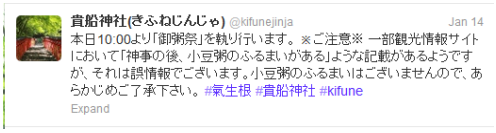

図 2: Twitter 上でのイベント予告

表 1: 使用した日時表現パターン

日付表現	yyyy 年 mm 月 dd 日 yyyy/mm/dd yyyy.mm.dd	mm 月 dd 日 mm/dd mm.dd
時刻表現	hh 時 nn 分	hh:nn
値の範囲	$2012 \leq yyyy$ $1 \leq dd \leq 31$ $0 \leq nn \leq 59$	$1 \leq mm \leq 12$ $0 \leq hh \leq 23$

そこで本節では，ソーシャルメディアからの予告型地域イベント自動収集手法，ツイートからのイベント情報抽出手法を提案する．また提案手法の性能について評価実験を行う．本稿では特に Twitter 上でのイベント予告情報を対象とする．

3.2.1 ソーシャルメディアからのイベント予告ツイート収集手法の提案

Twitter からイベント予告情報を収集する手法を提案する．

前述でも議論したとおり，イベント三要素が共通であれば，同じイベントが同定できると考えられる．そこで，本稿ではイベント三要素が含まれているツイートをイベント予告ツイートと定義する．イベント予告ツイートを収集するためには，イベント三要素を手掛かりとして用いる必要がある．そのうち，パターンが有限である「開催場所」と「開催日時」を手掛かりとして用いた．「開催場所」については対象を日本国内に限定した上で，都道府県名及び市町村名を用いる．「開催日時」については，表 1 にある日時表現パターンを「開催日時」の手掛かりとして用い，収集時点よりも未来の日時になるように制約を加えることとする．

3.2.2 イベント予告ツイートからのイベント情報抽出

ここでは，ソーシャルメディアから収集したイベント予告ツイートからイベント情報を抽出する手法について述べる．

収集したイベント予告ツイートは各投稿者により自由に記述されており．決まったフォーマットは無い．そのため，ツイートからイベント情報を抽出するためには，自然言語処理の技術を用いて，柔軟にイベント三要素を抽出する手法が必要となる．本稿では，ヒューリスティックスなパターンマッチング及び辞書を用いて，各要素を抽出する手法を提案する．

イベント名 イベント名の抽出には 3 つの手掛かりを用いる．

- **括弧** 括弧に囲まれた文字列をイベント名として抽出する．括弧として「」【】＜＞ () [] を用いる．
- **手掛かり表現** 特定の手掛かり表現の前に出現する語をイベント名として抽出する．実際に「〜が開催される」「にて開催される〜」「〜が現在開催中」を用いた．
- **末尾表現** 特定の末尾表現を含む文字列をイベント名として抽出する．実際には，るるぶ.com から収集したイベント名の頻出末尾表現を用いた．

開催場所 開催場所の抽出には 3 つの手掛かりを用いる．

*1 http://www.rurubu.com/event/
*2 http://zaq.ne.jp/event/

表 2: イベント情報抽出手法 評価実験

評価対象 /手法)	精度	再現率	F 値
告知ツイート収集	0.745 (114/153)	1.00 (114/114)	0.854 -
イベント名抽出	0.736 (39/53)	0.650 (39/60)	0.690 -
開催場所抽出 （提案手法）	0.609 (92/151)	0.829 (92/111)	0.702 -
開催場所抽出 （従来手法 [遼平 08]）	0.720 (77/107)	0.694 (77/111)	0.710 -

地名辞書 出現する地名を開催場所として抽出する．

手掛かり表現 イベント名と同様に手掛かり表現を用いる．

末尾表現 イベント名と同様に末尾表現を用いる．

開催日時 開催日時抽出には，表 1 の表現パターンを用いた．

3.2.3 データ収集及び精度評価

本節では，提案する予告イベントツイート収集手法及びイベント情報抽出手法について評価実験を行った．

日付表現として 8 月 1 日から 8 月 7 日，地名として 47 都道府県名，区市町村名を用いて予告イベントツイートの収集を試みた．予告イベントツイートは全部で 7649 件得られた．そのうち，153 ツイート（全体の 2%）を抽出し評価を行った．この結果を表 2 に示す．評価結果より，精度 0.745，再現率 1.00 で予告イベントツイートを収集することができた．

次に，抽出した 153 ツイートにイベント情報抽出手法を適用した．また地名抽出の従来手法として笹野らの手法を適用した [遼平 08]．実験結果は，同じく表 2 に示す．イベント名については，精度 0.736，再現率 0.650 で抽出することができた．また表 2 より，括弧，末尾表現が，イベント名抽出の手がかりとして有効であることがわかる．開催場所については，提案手法を用いて精度 0.609，再現率 0.829 で抽出することができた．また表 2 より，いずれの手掛かりも地名抽出に有効であることがわかる．また，従来手法では精度 0.720，再現率 0.694 で地名抽出を行うことができた．

提案手法と従来手法の地名抽出性能を比較すると，従来手法は前後の文脈から地名と判定可能な語を出力するため精度は高い．しかし,「8.4 土 @立川」「 8 月 4 日 土 京橋」のように学習コーパス（新聞記事）に出現しない文脈による地名の判定は難しい．それに対し，提案手法は地名辞書にある語を全て地名として抽出するため，精度は低く，再現率が高い．

実験結果より，予告イベントツイートの収集については高い精度，再現率で実施することができた．また，イベント情報の抽出についても，実用に近い精度で抽出することができた．今後は，ソーシャルメディア上のテキストをコーパスとして地名抽出の手掛かりとなる文脈情報を学習することで，精度向上を目指すことを考えていきたい．

4. ユーザのイベント参加状態推定

ここでは，予告型地域イベントを収集するための手法について説明する．

ユーザのイベント参加状態を推定する手法は 2 つのアプローチが考えられる．一つは，参加状態を知りたいユーザを常時監視するアプローチである．このアプローチでは，対象となるユーザのイベント参加状態が網羅的に収集できる反面，大量のユーザを対象とすることは，実現制約上，困難である．もう一つは，あるイベントに関する情報を監視することで，参加しているユーザを抽出するアプローチである．このアプローチは，監視する情報を適切に設定する必要がある反面，情報検索を手法を用いる事で大量のユーザを対象にすることが可能である．

本稿では，大量のユーザを対象とするために，イベント情報を監視することで参加ユーザを自動抽出する手法を提案する．なお，ユーザの参加状態推定そのものについては，今後の課題とする．

4.1 ユーザのイベント参加状態推定のための参加ユーザ自動抽出手法

ここでは，イベント情報を手がかりとしたイベント参加ユーザの自動抽出手法について説明する．実際には，対象イベントの三要素を入力とし，そのイベントへの参加ユーザを抽出する手法である．本手法では，下記の 2 つの段階により，当該イベントに参加してるユーザ集合を得る．

STEP 1 情報検索の手法を用いて,「イベント名」「開催場所」を言及しているユーザ候補を絞り込む．

STEP 2 得られたユーザ候補のうち，言及している時刻が「開催日時」の時間帯であるユーザを参加ユーザ集合とする．

STEP 1 ではイベント参加ユーザの絞り込みを行う．実際には，Twitter 検索 API において「イベント名」「開催場所」を検索クエリとして用いることで，イベント参加ユーザによると推測されるツイートを収集する．STEP 1 において，ユーザが「イベント名」や「開催場所」について言及する場合，表記の曖昧性のために，同じ「イベント名」や「開催場所」が様々な表記で投稿されると考えられる．しかし，通常のキーワード検索においてはこのような曖昧検索の機能は実現されていない．そのため，適切な検索クエリを設定する必要がある．適切な検索クエリを設定する手法については，下記で詳細に述べる．

STEP 2 においては，収集したツイートのうち，開催日時の時間帯及びその前数時間に投稿されたツイートのみを抽出する．それらを投稿したユーザをイベント参加ユーザとみなす．

STEP 1 において、キーワード検索を用いてイベントに参加しているユーザを発見するためには，検索クエリを適切に設定する必要がある．実際には，下記の 2 段階により行う．

STEP 1-a 情報検索の手掛かりとする「イベント名」「開催場所」について検索クエリ候補を作成する

STEP 1-b 各検索クエリ候補について，検索クエリとしての適切性を判定する

STEP 1-a において,「イベント名」「開催場所」を表す文字列の部分文字列を検索クエリ候補とする．

STEP 1-b においては，検索結果の空間局在性という指標を定義し，それを用いて検索クエリとしての適切性を判断する．検索結果の空間局在性とは，ある検索クエリを用いて検索した場合に得られた検索結果に紐付いた地名が，どの程度偏っているかを表す指標である．ある地域で行われるイベントについては，その地域の周囲に住んでいる人により言及される可能性が高いと考えられる．そのため，あるキーワードが地域イベントを一意に意味する場合，そのキーワードを投稿しているユーザの居住地は特定の地域に偏って存在していると思われる．つまり，イベントを表す検索クエリが適切である場合，検索結果の空間局在性は高くなると考えられる．

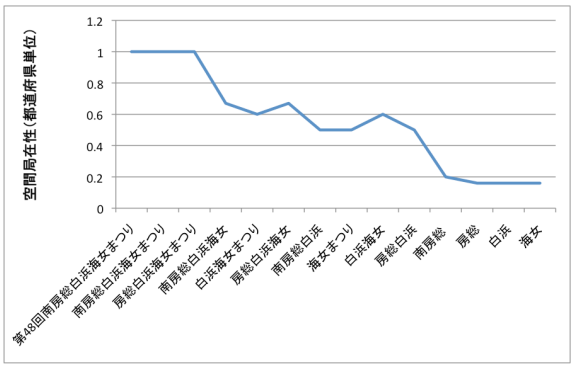

図 3: 「南房総白浜海女祭り」検索結果に占める千葉県在住ユーザの割合

表 3: ユーザイベント状態推定 評価実験

イベント名	開催場所	イベント名/開催場所
0.750 (36/48)	0.51 (30/59)	0.864 (51/59)

実際には，ある検索クエリに対する Twitter 検索結果に対し，各ツイート投稿者の居住地を判定し，集計する．そして，最も投稿者が多い都道府県とイベント開催場所の都道府県が一致していた場合に，そのクエリが適切であると判定する．投稿者の居住地は，Twitter ユーザプロフィールの「Location」の記述から判定した．

図 3 は「南房総白浜海女祭り」を検索クエリとして得られたツイートについて全投稿者に対し，開催場所の都道府県に在住するユーザの割合を表したグラフである．実際に「南房総白浜海女祭り」は千葉県で開催されるイベントであったため，各検索クエリによる検索結果に占める千葉県在住者の割合を表している．グラフより,「南房総白浜海女祭り」「南房総白浜海女」「南房総白浜」「房総白浜海女祭り」「房総白浜海女」「房総白浜」「白浜海女祭り」「白浜海女」「海女祭り」において千葉県在住者の割合は 50% を超えており，これらは適切な検索クエリと判定される．

4.2 参加ユーザ自動抽出手法の評価実験

本節では，提案手法の評価実験を行う．評価実験においては，提案手法により作成された検索クエリにより実際にイベント参加ユーザ集合が得られるかどうかにより評価を行った．

評価対象イベントとし，るるぶ.com から収集したイベントのうち，2012 年 7 月 19 日〜21 日の 3 日間に開催された 59 イベントを対象とした．評価結果は表 3 に示す．

表 3 より，59 イベントのうち，イベント名から適切な検索クエリが作成可能なイベントは 48 件 で，そのうち 75% のイベントについて参加ユーザによると思われる投稿を収集することができた．また，表 3 より，59 イベントのうち，開催場所名から適切な検索クエリが作成可能なイベントは 59 件 であり，そのうち 30% のイベントについて適切な検索クエリを作成し，参加ユーザによると思われる投稿を収集することができた．

これより提案手法をイベント名，開催場所名両方に適用することにより，51 件，86%のイベントに有効な検索クエリが作成できた．

これより，イベント名から適切なクエリを自動生成する場合に提案手法を適用することで，実用的な精度が得られることがわかった．一方，地名から適切なクエリを自動生成する場合には精度が低かった．しかし，両者により生成された検索クエリ用いることで，高い精度でイベントに参加しているユーザを収集可能であると言える．

5. おわりに

本研究では，ユーザへの行き先推薦のための「予告型地域イベントの情報収集システム」を構築することを目指し，基礎的な手法の提案を行った．

まず，予告型地域イベント情報の収集手法について，1. イベント情報サイトからの収集，2. ユーザ発信情報からの収集，という 2 つの手法を提案した．評価実験の結果，イベント予告ツイート収集手法は実用的な精度で収集を行うことができた．一方，ツイートからのイベント名抽出，地名抽出の精度は一定の精度はあるものの，今後，固有表現抽出や構文解析などによる精度の向上が必要であると考えられる．

次に，ユーザのイベント参加状態推定のために参加ユーザの自動抽出手法を提案した．実際には,「開催場所」「イベント名」を手掛かりとし，さらに検索クエリの適切性の指標としては，空間的な局在性を用いた．評価実験の結果，イベント名，開催場所名の両方から検索クエリを生成することで，実用的な精度でイベント参加状態のユーザを抽出することができた．今後は，文脈情報を用いてユーザ毎のリアルタイムなイベント参加状態を推定することで，大規模かつ高精度なイベント参加ユーザの抽出を実現していきたい．

参考文献

[Aramaki 11] Aramaki, E., Masukawa, S., and Morita, M.: Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter, in *Proceedings of the 2011 EMNLP*, pp. 1568–1576 (2011)

[Chakrabarti 11] Chakrabarti, D. and Punera, K.: Event Summarization Using Tweets, in *Proceedings of the ICWSM 2011)*, pp. 66–73, Barcelona, Spain (2011), AAAI Publications

[Hong 12] Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., and Tsioutsiouliklis, K.: Discovering geographical topics in the twitter stream, in *Proceedings of the WWW 2012*, pp. 769–778, New York, NY, USA (2012), ACM

[Lee 11] Lee, C.-H., Wu, C.-H., and Chien, T.-F.: BursT: a dynamic term weighting scheme for mining microblogging messages, in *Proceedings of the ISNN 2011*, pp. 548–557, Berlin, Heidelberg (2011), Springer-Verlag

[Sak 10] Earthquake shakes Twitter users: real-time event detection by social sensors, *Proceedings of the WWW 2010*, pp. 851–860 (2010)

[Watanabe 11] Watanabe, K., Ochi, M., Okabe, M., and Onai, R.: Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs, in *Proceedings of the CIKM 2011*, pp. 2541–2544, New York, NY, USA (2011), ACM

[笹野 08] 笹野 遼平, 黒橋 禎夫：大域的情報を用いた日本語固有表現認識, 情報処理学会論文誌, Vol. 49, No. 11, pp. 3765–3776 (2008)