情報処理学会論文誌 Vol.57 No.1 123–132 (Jan. 2016)

マイクロブログを用いたイベント情報抽出技術

山田 渉[1,a)] 菊地 悠[1] 落合 桂一[1] 鳥居 大祐[1] 稲村 浩[1] 太田 賢[1]

受付日 2015年4月9日, 採録日 2015年10月2日

概要：本論文では，大量かつ広範な話題が投稿されるマイクロブログの *Twitter* からイベント情報を自動的に抽出する技術を提案する．従来，イベントの自動抽出にはジオタグが付与された投稿件数の急上昇を検知することでイベントを発見する手法が用いられていた．しかし，この手法で検知できるイベントは開催中のイベントに限られており，イベントの名称や開催期間等の詳細情報や，開催中のものだけでなく将来に開催されるイベント情報は抽出できなかった．そこで本研究では *Twitter* におけるイベントの告知に関する投稿に着目し，機械学習を利用してイベントの名称，開催場所，開催期間の情報を抽出する．本手法は次の 4 段階の処理で構成されている．(1) あらかじめ用意した地名のリストを用いて，ツイートと呼ばれるユーザの投稿文と地名を関連付ける．(2) 地名と関連付けられたツイートの中から，Support Vector Machine を用いてイベントの告知に関するツイートを抽出する．(3) Conditional Random Fields を用いてイベントの名称と開催期間を抽出する．(4) イベント名称の類似度と開催場所を用いて，表記揺れのある重複したイベント情報の名寄せをする．提案手法の有効性を確認するため，提案手法の再現率と適合率およびイベントの開催期間に対する抽出日について評価をした．その結果，従来法と比較して高い適合率でイベント情報が抽出可能なうえ，現在開催中のイベントだけなく将来のイベント情報も抽出可能なことを確認した．

キーワード：イベント情報抽出，Twitter，マイクロブログ，自然言語処理，機械学習

Extracting Local Event Information from Micro-blogs

Wataru Yamada[1,a)] Haruka Kikuchi[1] Keiichi Ochiai[1]
Daisuke Torii[1] Hiroshi Inamura[1] Ken Ohta[1]

Received: April 9, 2015, Accepted: October 2, 2015

Abstract: This paper describes a method to extract local event information from the micro-blog service Twitter that holds innumerable user-posted short messages. Previous methods detect event by using surge of geo-tagged user-posted message. However, the approach is only able to detect that event occurs and impossible to extract future event information or detailed event information such as name and holding time. This paper extract event name, venues and holding time of event information from tweets related to local event using machine learning. Our approach composed of four steps: (1) relate user-posted message called tweet to venue by using list of venue, longitude and longitude, (2) extract tweets related to local events from local tweets by Support Vector Machine (SVM) approach, (3) identify and extract names and times of local event from tweets related to local event by applying Conditional Random Fields (CRF), (4) aggregate duplicated local event information by using venues and similarity of names. We implemented the proposed method and evaluate it. As a result, we confirmed that it can extract not only local event information in session but also future one with higher precision than conventional method.

Keywords: event information extraction, Twitter, micro-blogs, natural laungage processing, machine learning

1. はじめに

近年，施設情報，道路状況やイベント情報といった地理空間情報が多数の位置情報サービスによって提供されてい

[1] 株式会社 NTT ドコモ
NTT DOCOMO, INC., Yokosuka, Kanagawa 239–8536, Japan
[a)] wataru.yamada.rz@nttdocomo.com

る．たとえば *Trip Advisor* [*1]は 3,200 万件超ものホテルやレストラン情報やその評判情報を提供し，ぐるたび [*2]は日本各地の名物グルメ情報を提供している．これらの位置情報サービスは，ユーザの観光やお出かけにおけるプランニングにおいて有用である．また地理空間情報は，位置情報サービスのような消費者向けのサービスだけでなく，都市計画や防災計画等においても重要であり，その活用が進められている [1]．

しかし，有益な地理空間情報の活用のためには，情報の正確性だけでなく，件数や詳細さといった情報の量，情報の更新頻度といった観点からも優れた地理空間情報が重要である．たとえば位置情報サービスを例にした場合，情報の件数が少なく興味のある情報が見つけられない場合や，更新頻度が少なく新たなスポットやイベント，新商品が提供されない場合，サービスの有用性は低下する．

しかし，地理空間情報の量と品質を維持するためには多大なコストがかかる．さらに地理空間情報の中でもイベント情報は一時性が高く動的であるため，情報の鮮度を保つためには頻繁な更新が必要となり，手動での更新には限界がある．

そこで本研究では大量かつ広範な非構造化データを持つマイクロブログの 1 つである *Twitter* [*3]に着目し，イベントの情報（イベント名称・開催場所・開催期間の 3 つ組）を自動かつ高精度に抽出する手法について検討する．イベントの情報源として，本研究で *Twitter* に着目した理由としては，*Twitter* を情報源として活用することで，多様かつ大量のイベント情報の抽出が期待できるためである．*Twitter* や *Facebook* [*4]等 Social Network Service（以下 SNS）の普及とともに，SNS 上で，日々の出来事やニュース，新商品やイベント情報といった様々な情報が共有されるようになった．SNS を用いたイベント情報の告知は，誰でも手軽に行うことができるため，自治体の観光課等のサイトでも告知がされる花火大会や地域のお祭りといった公共性の高いイベントだけに限らず，店舗のフェアやインディーズバンドのライブといった様々な種類のイベント情報が告知されている．

本研究のように SNS 上のユーザの投稿データを解析し，実世界で起きている様々なイベントを自動的に検出することは広く検討されてきた [2], [3], [4]．これらの研究の多くが検出できる事象は，各地で開催されている大規模なイベントの有無に限られていた．

イベントの有無だけでなく，イベントの名称や開催期間もあわせて抽出することが可能になれば，たとえば位置情報サービスに活用した際のユーザの利便性をおおいに向上させることや，イベント名称を含んだツイートからイベントの評判を推定する等，より詳細なイベント情報の分析が可能となる．さらに開催中のイベントだけでなく将来のイベント情報や，小～中規模なイベント情報も抽出することが可能になれば，より多くの活用が可能となる．

イベントの有無以上のイベントに関連した情報を抽出可能な研究としては，渡辺らの研究 [5] がある．渡辺らの研究は，ツイートからイベントに関連したキーワードを抽出することを目的としているが，適合率は 47%にとどまる．

そこで本研究では，イベントの名称と開催期間，開催場所の 3 つ組を，従来よりも高い適合率で各ツイート文中から抽出可能なシステムを提案する．高い適合率でイベント情報の 3 つ組の抽出を実現するために，本研究では従来研究 [2], [3], [4] のようなユーザの投稿データの地域的・時間的な件数の変化に着目するのではなく，イベントの告知ツイートとイベント名称が持つ自然言語的な特徴に着目した．そしてイベントの告知ツイートとイベント名称を機械学習によって抽出を行うことで，高い適合率で開催場所とイベントの有無という情報だけでなく，イベント名称や開催期間といったより詳細な抽出できることを明らかにする．

さらに本研究では，抽出したイベント名称を名寄せする機能についても検討を行う．ツイート中では短縮名称や表記揺れを含む名称でイベントが表記される場合がある．たとえば，“ROCK IN JAPAN FESTIVAL 2014” という正式なイベント名称に対してユーザによって “ROCK IN JAPAN” や “ROCK IN JAPAN 2014” といった異なる表記が用いられるという問題である．異表記の同一イベントは，それぞれを別のイベントとしてユーザに提示するよりも 1 つのイベント名称に集約するほうが望ましいため，本研究ではイベント情報の名寄せ処理についても検討を行う．

提案手法は 4 つのステップで構成されている．まずあらかじめ用意した地名リストを用いて，地名とツイートを関連付ける．次に地名と関連付けたツイートからイベントに関連したツイートを Support Vector Machine（以下 SVM と記載）[6] によって抽出をする．次に Conditional Random Fields（条件付き確率場，以下 CRF と記載）[7] によって，イベントの名称と開催期間を抽出する．最後にイベント名称の類似度と開催場所を用いて，名寄せ処理を行う．

提案手法における適合率と抽出件数を，地名と関連付けられた約 2,363 万件の 2013 年内に投稿されたツイートを用いて評価を行った．その結果，9,781 件のイベント情報が抽出された．さらに抽出されたイベント情報から 400 件を無作為抽出し，適合率を評価したところ，69%のイベント名称が正しく抽出され，従来手法 [5] の適合率を上回った．

2. 関連研究

ブログやマイクロブログ等のユーザ投稿型の Web サービスを情報源として，イベント情報を検出する研究は多数

[*1] Trip Advisor, http://www.tripavisor.jp/
[*2] ぐるたび，http://gurutabi.gtavi.co.jp/
[*3] Twitter, https://twitter.com
[*4] Facebook, https://www.facebook.com/

行われてきた．イベント情報を検出するための代表的な手法として，測位された位置情報付きの投稿の地域ごとの件数の急上昇を検出する手法がある [2], [3], [4], [8], [9]．これらの手法では，主に地震の発生や有名アーティストのライブ等の実時間で起きているイベント情報を対象としており，イベントが発生する前に情報を抽出することはできない．さらに抽出可能なイベント情報は，位置情報付きの投稿の地域ごとの投稿件数が大きく変化するほどの大規模なものに限られる．

一方，本研究のように投稿文書を解析することによって，イベント等の地域情報を抽出する手法も提案されている．岡本ら [10] は地名を本文に含んだブログ記事を収集し，地域性と時事性の高い話題のキーワードをクラスタリングし，地域イベントの抽出を行っている．また渡辺ら [5] は本文に地名を含むツイートの投稿時刻と内容の類似性からイベント情報の抽出を行っている．イベントと関連したキーワード群の抽出を目的としており，いずれの研究も開催中のイベント情報だけでなく将来のイベント情報も抽出可能である．しかし，イベントの名称や開催期間を特定して抽出することはできない．さらに，いずれの研究も位置情報付きの投稿の地域ごとの件数の急上昇を監視する方法のように，普段よりも多くの投稿がされている地域を対象として抽出するため，少数の投稿のみに記載される小規模なイベント情報を抽出することはできない．

本研究は，クラスタリング等の複数ツイートを必要とするアプローチと異なり，イベントの告知に関するツイートを対象にして，各ツイートに含まれるイベントの名称，開催期間，開催場所の 3 つの情報を特定して抽出をする．提案手法では，イベントの告知に関するツイートが 1 件でも存在すれば抽出対象となるため，開催中の大規模なイベントだけでなく，将来の小規模なイベントから大規模なイベントまでを抽出の対象とすることが可能である．さらに提案手法では，イベントに関連したキーワード群を対象とする [2] のではなく，イベントの名称・開催期間・開催場所の 3 つの情報を特定して抽出可能な特徴がある．

一円ら [11] は異なる位置情報サービスで提供されている POI（Point of Interest，以下 POI と記載）の情報を，POI の住所・電話番号・名称の編集距離と最長共通文字列長を用いて名寄せをする処理を提案している．また荒川ら [12] は Flickr の画像群をクラスタリングし，クラスタが示す最も確からしい POI を，名称の Jaro-Winkler 距離 [13] と単語の出現頻度を用いて推定する手法を提案している．本研究のイベント情報の名寄せ処理は，これらの研究と同様に名称の類似度を用いる．なお，“桜祭り” や “椿祭り” 等のような，同一の名称のイベントが全国各地で同時期に開催される場合があるため，名寄せ処理は開催場所ごとに行う．

3. イベント情報の自動抽出の要件

本研究では各ツイートに含まれるイベント情報を自然言語処理により抽出する．イベント情報はイベントの有無だけでなくイベント名称や開催期間，開催場所といった情報を含むものとする．ただし，開催期間については，ツイート文中の URL 先の Web サイトにのみ記載されている場合や，イベントの開始日や終了日のいずれかのみが記載されている場合があるが，本研究ではツイート文中に記載されている開催期間のみを対象として抽出を行う．

イベント名称は，短縮名称や表記揺れのある名称でツイートされる場合がある．異表記の同一イベントは，それぞれを別のイベントとしてユーザに提示するよりも，1 つのイベントにまとめたほうが望ましいため，本研究ではイベント名称を集約する名寄せ処理についても検討を行う．

以上から本研究が満たすべき要件は次の 2 点である．

1. イベント名称および開催場所をツイートから自動で抽出すること．ツイート文中に開催期間が記載されている場合，その開催期間も抽出すること．
2. 抽出したイベント情報の中で同一のものを自動で集約すること

4. 提案手法

本研究では，日本語ツイートから地名に関連するツイートを抽出し，さらにイベント情報（イベント名称・開催期間・開催場所の 3 つ組）を SVM と CRF を用いて段階的に抽出する手法を提案する．

4.1 地名抽出部

提案手法では**図 1** のように，まず地名抽出部を用いて，地名とツイートの関連付けを行う．地名とツイートの関連付けは図 1 の①から③までの 3 つのステップで構成されている．

第 1 に日本語のツイートを対象として，形態素解析を実施する（図 1-①）．次にあらかじめ用意した地域や施設の名称と後述する曖昧性を示すフラグを含む地名リストを参照し，地名と一致する名詞を本文中に含むツイートを抽出する（図 1-②）．最後に曖昧性がある地名を含むツイートに対してフィルタリングを行う（図 1-③）．曖昧性のある地名とは，人名や同名で異なる場所が存在するため曖昧性があり，特定の場所のことを示すとは限らない地名のことである．たとえば，人名と地名の曖昧性の例としては，苗字の “松島” と宮城県の観光名所の “松島” 等がある．同名の地名が存在する例としては，京都府の “円山公園” と北海道の “円山公園” 等があげられる．

同名の地名どうしの曖昧性については，落合ら [14] と同様に共起語を用いて曖昧性を除去する．本論文の共起語とは，各地名と共起しやすい語であり，近隣の地名や，その

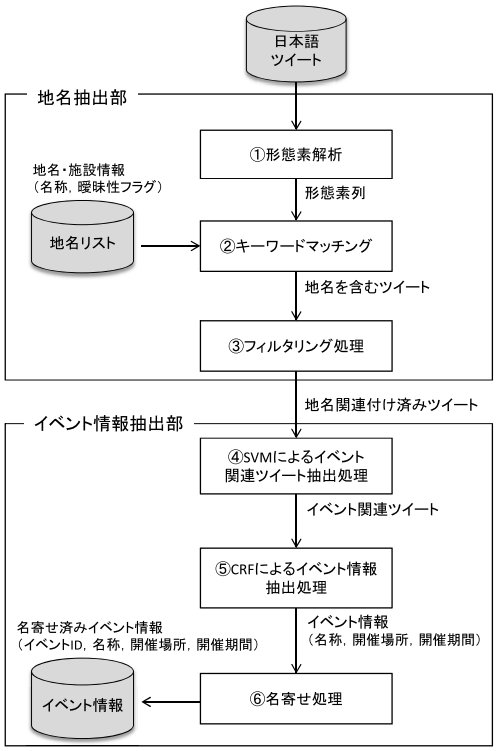


図 1 提案処理概要
Fig. 1 Procedure of proposed method.

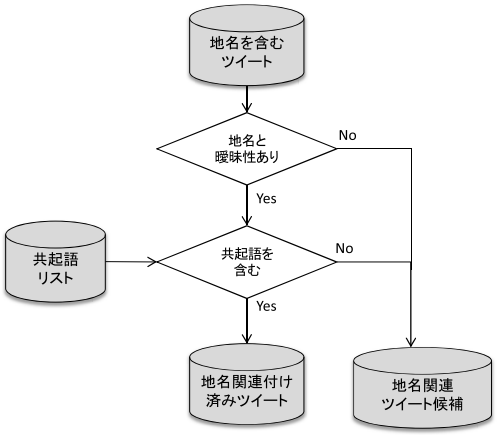


図 2 同名の地名との曖昧性除去
Fig. 2 Co-occurrence place-name identification flowchart.

地域特有の語等が該当する．**図 2** のようにツイートに含まれる地名が，同名の地名との曖昧性がある地名であった場合には，その地名だけでなくさらに共起語を含むツイートのみを該地名と関連するものとして抽出を行う．人名等と重複している地名については**図 3** のように共起語と CRF を併用して曖昧性を除去する．人名の曖昧性の除去に CRF を利用する理由は，共起語による方法のみを用いた場合よりも多くのツイートを地名と関連付けるためである．CRF を用いた曖昧性の除去は地名と人名のような文章中での使われ方が異なるもののみに適用が可能なため，同名の地名の曖昧性の除去に関しては共起語による方法のみを用いている．

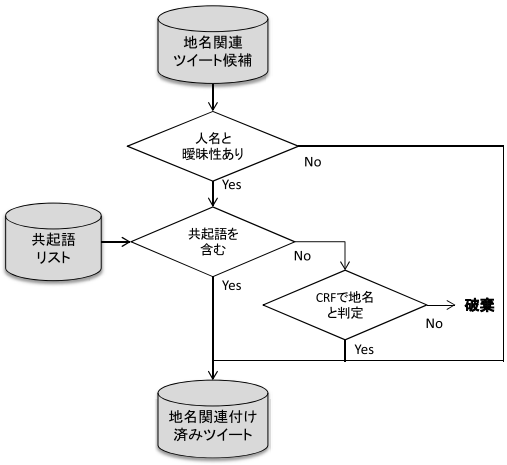


図 3 人名との曖昧性の除去
Fig. 3 Person-name and place-name identification flowchart.

以上の処理によって，地名抽出部は曖昧性を除去して日本語ツイートを地名と関連付けてイベント情報抽出部へと出力をする．

イベント情報抽出部では，地名関連付け済みツイートから，イベント名称と開催期間の抽出を行う．本処理は大きく 3 つのステップで構成される．第 1 に，地名と関連付け済みのツイートからイベント名称や開催期間といった情報を含むツイートを抽出するイベント関連ツイート抽出処理を行う（図 1-④）．第 2 に，イベント情報を含むツイートから，イベント名称，開催期間の抽出を行う．開催場所の情報には地名抽出部を用いてツイートと関連付けられた地名を割り当てる（図 1-⑤）．第 3 に，抽出されたイベント情報に対して，開催場所とイベント名称の類似度を用いて同一のものかを判定する名寄せ処理を行う（図 1-⑥）．また名寄せ処理では同一と判定されたイベント情報には同一の ID を割り当てる．異なると判定されたイベント情報には異なる ID を割り当てる．

4.2 SVM によるイベント関連ツイート抽出

地名と関連付け済みのツイートの中からイベント情報を含むツイートをフィルタリングする処理では機械学習を用いる．機械学習を用いないアプローチとしては，あらかじめ正規表現 [15] 等で単語のパターンを定義し，抽出する方法がある．しかし，多くの件数および種類のイベントに関連したツイートを抽出する無矛盾なルールを人手で構築するのは困難であるため，本研究では教師あり機械学習の一種である SVM を用いる．SVM を用いる理由として，提案手法ではツイート文中の単語を特徴として用いるため特徴の次元数が大きくなるが，SVM は従来からある学習モデルと比べて汎化性能が高いうえ，高次元の特徴を用いても過学習しにくいという特長があることがあげられる．**図 4** のように，SVM によるイベント関連ツイート抽出処理は学習フェーズと推定フェーズに分かれる．まず学習フェー

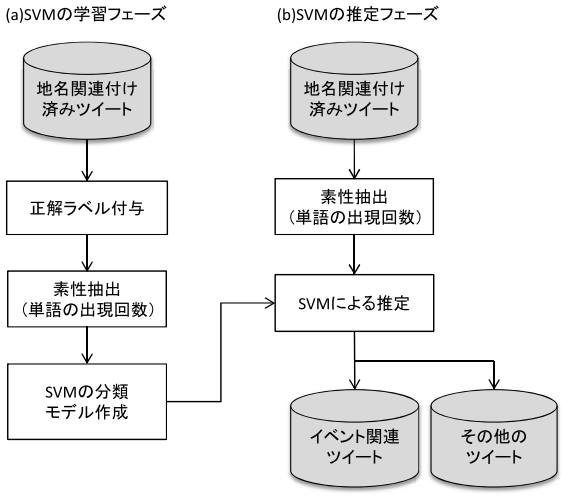


図 4 SVM によるイベント関連ツイート抽出処理
Fig. 4 Procedure of extracting event tweets by SVM.

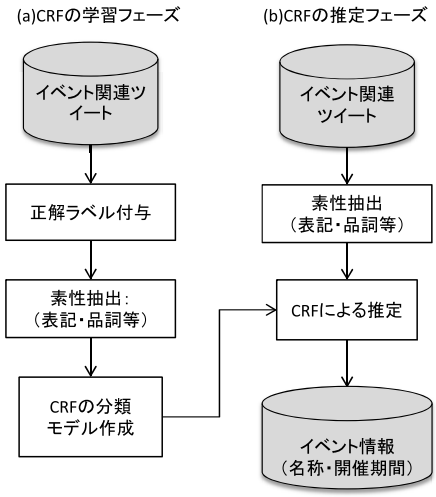


図 5 イベント情報抽出処理
Fig. 5 Procedure of extracting event information.

表 1 SVM の学習データの例
Table 1 Example of training data for SVM.

No.	正解ラベル	地名関連付け済みツイート
1	○	○○未来館で「21世紀の未来展」を12月3日から開催します.
2	○	○○未来館の21世紀の未来展に参加したいなぁ.
3	×	○○未来館はバリアフリー化活動を推進しています.
4	×	××で18時から飲み会を開催します.

ズでは SVM の学習データとして，**表 1** のようにイベントに関連しているかどうかの正解ラベルを付与した地名関連付け済みのツイートを用意する．またイベントと関連しているかという判断は，ツイート本文中に固有のイベント名称を含むか否かで決定する．たとえば，“21 世紀の未来展”というイベントが 12 月 3 日に開催される場合に，表 1 の学習データのように，“21 世紀の未来展” という固有のイベント名称を含むツイートを正例としている．

次に正解ラベルを割り当てた地名関連付け済みツイートに対して，素性 [16] の抽出を行う．素性とは言語処理における特徴量のことで，提案手法では素性として文中に各単語が何回出現したかという出現回数を用いている．そして抽出した素性を用いて SVM を学習させることで，各単語の出現回数ごとの重みが計算され，分類モデルが作成される．推定フェーズでは，まず地名関連付け済みの各ツイートの学習データに含まれる単語ごとに出現回数が計算される．そして，学習フェーズで構築した分類モデルを用い，地名関連付け済みの各ツイートがイベント情報を含むかどうかを判別する．たとえば，SVM のカーネル関数として線形カーネルを用いた場合，各単語の出現回数とその重みの内積によって判別される．イベント情報を含むと判別されたツイートは CRF によるイベント情報抽出へと出力される．

4.3 CRF によるイベント情報抽出

イベント情報抽出処理は，SVM によってイベント固有の名称を含むと判別されたツイートを対象とし，本文に含まれるイベント名称を，また可能であれば開催期間を抽出する．複数の単語からなる意味的あるいは文法的なまとまりを抽出することを言語処理では固有表現抽出 [16] と呼ぶ．本研究はイベント名称と開催期間という 2 つの固有表現を抽出することに相当する．イベント名称と開催期間情報の固有表現抽出には，教師あり機械学習の一種である CRF を用いる．CRF は入力された要素が連なった系列に対し，その素性に基づいてあらかじめ定められたラベルを付与する系列ラベリング問題向けの学習器である．ここでは文書を構成するそれぞれの語に対して，イベント名称や開催期間の固有表現の一部，またはそれ以外であることを表すラベルを付与する．CRF は同様の教師あり機械学習の手法である隠れマルコフモデル（Hidden Markov Model，HMM）と異なり，文字の種類や品詞，文字数等の素性を定義して柔軟にモデルに組み込むことができる．たとえばイベントの名称には「祭り」や「ライブ」等の語あるいは「Live」のようなアルファベットが，開催期間には数字が含まれることが多い，といった複数の種類の特徴を考慮することが可能である．

CRF によるイベント情報抽出処理も，SVM によるイベント関連ツイート抽出と同様，**図 5** のように学習フェーズと推定フェーズに分かれる．学習フェーズでは，学習データとして**表 2** のような正解ラベルを付けたイベント関連ツイートを用意する．ラベルはイベント名称の固有表現の開始を表す B-Event ラベルとそれに続く固有表現の一部であることを表す I-Event ラベル，開催期間の固有表現の開始を表す B-Time ラベルとそれに続く I-Time ラベル，その他の要素であることを表す O ラベルの 5 種類を付与する．次にラベルを付与した学習データから素性の抽出を行う．素性の抽出には既存の形態素解析ツールである JTAG [17] を用いた．ここでは JTAG の出力の単語の表記，品詞，原

形，読み，グループの 5 種類の情報に加え，文字数，文字種の 2 種類の情報を基に素性を作成し，さらにそれぞれの素性の重みを計算してモデルを構築する．

推定フェーズでは，イベント関連ツイートを学習データと同様に素性に変換する．学習フェーズでは構築したモデルを用い，各素性とその重みから最も確率が高いラベルを各形態素に割り当て，イベント名称および開催期間を抽出する．またイベント名称は括弧等の余分な文字列が付随する場合があるため，正規化を行い，最終的なイベント名称とする．

4.4 最長共通部分列比による名寄せ処理

CRF により固有表現として抽出されたイベント名称は，ユーザによって表記が異なることがある．たとえば「21 世紀の未来展」と「21 世紀のみらい展」のように同一のイベントが複数の表記で抽出されることがある．そのため提案手法では，開催場所とイベント名称の類似度を用いてイベント情報の名寄せ処理を行う．

提案手法では，**図 6** のように抽出したイベント情報を開催場所ごとにグルーピングを行う．次に開催場所が同じイベント情報どうしのペアを全通り作成する．そして作成したペアどうしのイベント名称の類似度を計算し，閾値以上であれば，同一のイベント情報と判定し，それぞれに同一のイベント ID を割り当てる．

表 2 CRF の学習データの例
Table 2 Example of training data for CRF.

表記	品詞	正解ラベル
未来	名詞	B-EVENT
展	名詞接尾辞	I-EVENT
は	格助詞:連用	O
21	名詞:日時	B-TIME
日	名詞:日時	I-TIME
から	格助詞:連用	O

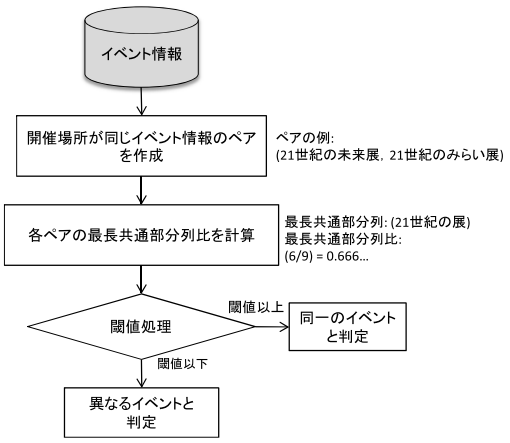


図 6 最長共通部分列比による名寄せ処理
Fig. 6 Procedure of aggregating event information using LCSR.

提案手法ではイベント名称の類似度に，最長共通部分列比（Longest Common Subsequence Ratio，LCSR）[18] を用いる．共通部分列とは，2 つの系列において，連続または非連続にかかわらず同じ要素が同じ順序で出現した部分列のことである．またとりうる共通部分列のうち，最も長いものを最長共通部分列（Longest Common Subsequence，LCS）といい，その長さを最長共通部分列長という．

提案手法では，式 (1) のように 2 つのイベント名称の最長共通部分列長を長い方のイベント名称の文字数で割った最長共通部分列比をあらかじめ設定した閾値と比較し，名寄せを行う．ただし，提案手法は略記や別記といった表記揺れには対応していない．これはたとえば，“オクトーバーフェスト” というイベント名を “OKF” や” ジャーマンフェス” といったように記載する種類のものである．これらの表記揺れの対応については今後の課題とする．

$$\mathrm{LCSR} = \frac{length(LCS(\boldsymbol{X}, \boldsymbol{Y}))}{\max(length(\boldsymbol{X}), length(\boldsymbol{Y}))} \quad (1)$$

5. 評価

提案手法によって抽出可能なイベント情報の適合率および再現率，データの傾向等を確認するため，実装を行い，**表 3** に示す実行環境で性能評価を行った．

5.1 処理ステップごとの評価

5.1.1 イベント情報抽出における評価

SVM および CRF によるイベント情報抽出の適合率および再現率を評価するため，SVM を用いず CRF だけでイベント情報抽出を行った場合と SVM と CRF をあわせてイベント情報抽出を行った場合の 2 通りで評価を実施した．評価で使用した SVM はカーネル関数に線形カーネルを使用したものである．

まずテストデータを用意するため，2013 年の地名関連付け済みのツイートを無作為サンプリングし，イベント情報を含むか含まないかを目視で確認した．その結果，**表 4** のように 1,888 件のイベント情報を含まないツイートと，112 件のイベント情報を含むツイートを得ることができた．

また SVM および CRF の学習データは同様に，2012 年の地名関連付けツイートから無作為サンプリングをし，目視でデータを確認することで作成した．SVM の学習デー

表 3 実験環境
Table 3 Experiment environment.

OS	Ubuntu/Linaro 4.6.3-1 ubuntu5
CPU	Intel® Xeon® CPU 5670 * 2
RAM	23GB

表 4 SVM によるイベント関連ツイートの分類結果
Table 4 Result of event tweet classification by SVM.

		正解ラベル		
		イベント関連	その他	合計
推定ラベル	イベント関連	31	44	75
	その他	81	1844	1925
	合計	112	1888	2000

表 5 イベント情報抽出の適合率および再現率
Table 5 Precision and recall of extracted event information.

	抽出件数	適合率	再現率
CRFの場合	64	50.0%	28.6%
SVM + CRF	19	78.9%	13.4%

タはイベント名称を含む正例のツイート 200 件およびイベント名称を含まない負例の 1,800 件のツイートの合計 2,000 件のツイートである．CRF の学習データはイベント情報を含むツイート 254 件である．

SVM によるイベント関連ツイートの抽出を行わずに，地名関連付け済みツイート 2,000 件に対して CRF のみでイベント情報抽出を行った場合，**表 5** のように 64 件のイベント情報が抽出された．しかし，抽出されたイベント情報のうち，人手で抽出したものと完全に一致したものは 32 件であり，適合率は 50%，再現率は 29%となった．一方，SVM によるイベント関連ツイート抽出処理を前処理として実施した場合，最終的に得られたイベント名称は 19 件で，適合率は 79%，再現率は 13%となった．このように SVM を適用することで，適合率が大幅に上昇することが確認された．このことから本研究が提案するイベント告知ツイートやイベント名称が持つ自然言語的な特徴を考慮し，段階的に機械学習でイベント情報を抽出することは，高い適合率でイベント情報を抽出する有効な手法といえる．また SVM の適用よって，最終的な適合率は上昇するが，再現率下がるため，実際に SVM を適用するかどうかは，抽出したイベント情報のユースケースによると考えられる．たとえば位置情報サービス等で，利用者に可能な限り誤検出した情報を提示しないというポリシの場合は，SVM と CRF をともに適用することが望ましい．一方で，各地域のイベント情報を可能な限り網羅する必要がある場合には，SVM によるイベント関連ツイートの抽出を適用せずに CRF によるイベント情報の抽出のみを実行することも考えられる．

5.1.2 名寄せ処理の評価

名寄せ処理における閾値ごとの正確度，適合率，再現率および F 値を評価した．評価データには，提案手法によって 2012 年 4 月 1 日のツイートから抽出された 55 種類の表記揺れをした同一のイベント名称のペアを正例として，714 種類の異なるイベント名称のペアを負例として用いた．その結果を**図 7** に示す．閾値を 0.38 および 0.39 に設定したとき，正確度および F 値が最も高くなり，正確度が 98%，F 値が 85%に達した．またこのとき適合率は 87%，再現率が 84%である．また適合率は閾値を上げるにつれ徐々に高くなる一方，再現率は徐々に低下した．

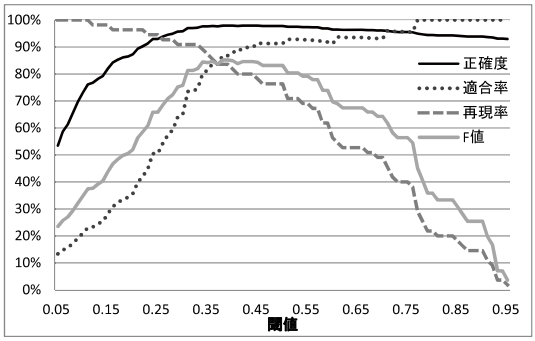


図 7 名寄せ処理における正確度，適合率，再現率および F 値
Fig. 7 Accuracy, precision, recall and F-measure of aggregating event information along LCSR.

名寄せの誤りの主な原因は，カナや漢字といったイベント名称の表記方法の違いである．たとえば，“ロックインフェスティバル” や “Rock in Festival” とペアの場合，最長共通文字列長は 0 である．このような場合，提案手法では，同一のイベント情報であっても異なるイベント情報と見なしてしまい名寄せできない．改善策としては，文字列の表記方法の揺れを考慮して，文字列の類似度に新たに単語の読み情報も加える方法が考えられる．

5.2 抽出件数および適合率の評価

提案手法は大量のイベント情報が，高い精度で抽出可能なことに加え，投稿された日よりも後に開催されるイベント情報も抽出可能であるように設計を行った．そのため，提案手法によるイベント情報抽出の抽出件数および適合率，抽出日に対するイベントの開催期間について評価を実施した．ただし，イベント情報の再現率については，日本全国のイベントの総数を測定することは困難であるため，Ling らの研究 [3] と同様に抽出件数を評価するのみにとどめる．

各ステップにおける学習データは 5.1 節と同一である．また名寄せ処理の閾値は適合率を重視し，5.1.2 項の評価実験において適合率が 90%でかつ再現率が 80%を下回らなかった 0.44 に設定した．評価データは 2013 年 11 月中に投稿された日本語ツイートを対象に，地名抽出部によって抽出された約 2,363 万件の地名関連付け済みツイートである．そこから SVM および CRF により，13 万件超のイベント名称を抽出し，さらに名寄せ処理をし，最終的に 9,781 件のユニークなイベント情報を得た．

しかし，抽出されたイベント情報の中には，2013 年 11 月以外に開催されるものや，誤抽出されたイベント情報と関係ないものが含まれている．そこで抽出したイベント情報から 400 件のイベント情報を無作為サンプリングし，サ

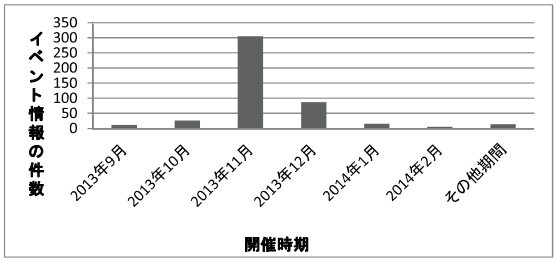


図 8 イベント情報の開催期間の分布
Fig. 8 Distribution of holding-time of event information.

表 6 開催期間と抽出日の関係
Table 6 Relationship between holding-time and extracted date.

	件数	割合
開催前	253件	63%
開催中	105件	26%
開催後	25件	6.3%
不明	17件	4.3%
合計	400件	100%

ンプリングしたイベントの開催時期を目視で調査をした．**図 8** は抽出した 400 件のイベント情報の候補の開催時期ごとの延べ件数であり，11 月を開催期間に含むものは，305 件存在し，全体 400 件の約 76%に相当する．

次に提案手法が開催前または開催中のイベント情報を抽出可能か調査するため，あるイベント情報が抽出された時期とそのイベントの開催期間について評価した．まず開催時期の評価と同様に，400 件のイベント情報を無作為サンプリングし，それぞれのイベント情報の開催期間に対して，イベント情報がいつ抽出されたかを調査した．**表 6** のように 253 件（63%）のイベント情報が開催期間よりも前に，105 件（26%）のイベント情報が開催期間中に抽出された．すでに終了した過去のイベント情報は，抽出されたイベント情報のうちの 25 件（6.3%）にすぎなかった．また残りの 17 件（4.3%）は，誤って抽出されたイベント情報か，開催期間が不明なものである．以上の結果から，提案手法は，現在の開催中のイベント情報が抽出できるだけでなく，将来開催されるイベント情報も抽出可能なことが確認された．

最後にサンプリングしたイベント情報 400 件を用いて，イベント名称の適合率の評価をするため，目視で以下の 3 種類に分類した．

1. 抽出したイベント名称が人手で抽出したものと完全一致するもの
2. 抽出したイベント名称が人手で抽出したものと一部一致するもの
3. イベント名称と関連のないもの

実際に抽出されたイベント情報の一部を**表 7** に示す．

表 7 抽出されたイベントの分類
Table 7 Classification of extracted event information.

No.	分類結果	抽出したイベント名称	ツイート本文
1	完全一致	アート手づくりフェスタ	11月30日(土)、東京国際フォーラムで 開催の「**アート手づくりフェスタ**」に出展します。アクセス：JR東京駅より徒歩10分、JR 有楽町駅より徒歩5分
2	部分一致	カピバラフェスタ2013の「大宮公園小動物園」	11/10(日)埼玉県こども動物自然公園で開催される**カピバラフェスタ2013**の「大宮公園小動物園」ブースに写真を展示していただけることに。5年前の2007年9月15日に長崎バイオパークで「ヤマト」君「コロ」ちゃん「彦馬」君の貴重な1枚 http://...
3	誤抽出	西武園ゆうえんち	RT：「西武園ゆうえんち」で毎年人気のイルミネーションが、今年も開催されます。テーマは「結婚したくなるイルミ」☆ 狩りの名手であるオリオンの「クラウン（王冠）」と、森の女神であるアルテミスの「ティアラ」が、冬の夜を煌びやかに輝かせます。https…

表 8 イベント情報の適合率の評価
Table 8 Precision rate of extracted event information.

	件数	割合
完全一致	276件	69%
部分一致	55件	14%
誤抽出	69件	17%
合計	400件	100%

表 9 抽出されたイベント情報の例
Table 9 Example of extacted event information.

No.	抽出されたイベント名称	抽出元ツイート	イベントの種類
1	花やしき de 年またぎカウントダウン2014	2014年は花やしきで迎える。浅草寺への初詣前に、2013年の遊び納め。『開園160周年記念「花やしき de 年またぎカウントダウン2014」』は2013年12月31日（火）浅草花やしきにて開催。http://...	レジャー施設のイベント
2	国際画像機器展2013	【 イベント情報 】「国際画像機器展2013」開催 ◆日程：12月4日～6日 ◆会場：パシフィコ横浜 ◆主催：日本画像・計測機器協議会 ⇒ http://..	大規模なビジネス向けの展示会
3	門司港グランマーケット2013秋	ようやく会場の設営もほぼ完了した感じです。明日は予定通り門司港グランマーケット2013秋開催となります。明日、歴史と古い町並みが調和する門司港でお会いしましょう。皆様、お気を付けてお越し下さい。http://...	地域のお祭り
4	両さんまつり	亀有で人気マンガこち亀「両さんまつり」開催 大人の社会見学ニュース 葛飾区亀有といえば週刊少年ジャンプで連載中の「こちら葛飾区亀有公園前派出所」の舞台として知られているが、マンガのゆかりの… http://...	地域のお祭り
5	中世の古文書	国立歴史民俗博物館に。「中世の古文書」という企画展。解説の時間に間に合って面白く聞けた。聞けば毎週土曜日の特定時間にしか聞けないそうでラッキー！一番おもしろいのは、天皇個人の財産は相続の時に税の代わりに国に物納されてここに。http://...	博物館の展示会
6	くまの親子とクリスマスケーキ	27日から中崎町のJAMPOTさんでクリスマスイベント「くまの親子とクリスマスケーキ」開催。くまの親子のお話に沿ったかわいい作品が沢山並びます！カシャロ作品も！ http://...	店舗のイベント
7	のんのんびより アニメ化フェア	【なんば店A】【フェア情報】「のんのんびより アニメ化フェア」を開催中！！とらのあなで対象商品お買い上げの方に先着で「特性キャラクターシール」をプレゼント！！詳しくはコチラへ⇒http://...	店舗のイベント
8	PAUL McCARTNEY OUT THERE JAPAN TOUR	チケットが高い……orz RT: ポール・マッカートニーは生で観る『PAUL McCARTNEY OUT THERE JAPAN TOUR』は2013年11月18日（月）～20日（水）東京ドームにて開催。	有名アーティストの音楽イベント
9	ネズミュージック2013秋	RT：まずはライブ告知。11/9(土) club edge 六本木にて開催される「ネズミュージック2013秋 夢見るくらいいいじゃない！」に出演します！上福岡雑業工房は13:30頃から演奏予定です。なんと入場無料！皆様ぜひお越しくださいませ！	音楽イベント

No.1 はツイート中に含まれているイベント名称が完全に抽出された例である．No.2 はイベント名称に加えて，他の文字列が付随している例であり，これらは部分一致と分類している．また No.3 は，誤抽出した例であり，イベント名称ではなく開催場所を抽出している．

以上のように抽出したイベント名称を分類した結果を，**表 8** に示す．渡辺らの研究 [5] ではツイッターからイベントと関連したキーワードを 47%の適合率で，岡本らの研

究 [10] ではブログから，ユーザが見てイベントと判断できるキーワード群を 39%の適合率で抽出可能なことが示されている．本研究で抽出したイベント名称をこれらのキーワードと見なして比較をすると，本研究は完全一致の場合に限っても 69%，部分一致しているものも含めると約 83%の適合率となり，従来研究を大きく上回った．

また**表 9** に抽出されたイベント情報の一部を示す．抽出されたイベントには，たとえばNo.1のような有名レジャー施設の大規模なイベント情報もあれば，No.2 のような大規模なビジネス向けの展示会等もある．No.1 や No.2 のような単一の施設のイベントが抽出されるとともに，No.3 や No.4 のような多数の出店が行われる地域のイベント情報といったものも抽出された．また No.5，6，7 に示すように博物館の展示会や店舗のフェア情報といった小規模なイベントも抽出された．さらに No.8 の有名アーティストの音楽イベントの情報が抽出される一方で，No.9 のようなインディーズバンドの音楽イベント情報といったものも抽出された．以上のように提案手法による抽出結果には様々な規模や種類のイベント情報が含まれていることを確認した．

6. 結論

本研究では大量かつ広範な話題が投稿されるマイクロブログの *Twitter* からイベント情報を自動的に抽出する技術を提案し，評価を行った．提案手法は SVM と CRF を用いて段階的にイベントの名称，開催場所，開催期間の 3 つ組のイベント情報を抽出する．イベント名称の類似度と開催場所を使ったイベント情報の名寄せ処理についても検討した．

本研究ではイベントの告知ツイートやイベント名称が持つ自然言語的な特徴に着目し，段階的に機械学習を用いてイベント情報を抽出する手法を提案した．そして提案手法が各地で開催されている大規模なイベントの有無を検出することまででであった既存研究 [2], [3] と異なり，将来のイベント情報まで対象を広げてイベントの詳細情報を抽出可能なことを示した．さらに提案手法がイベントに関連したキーワードを抽出可能な既存研究 [5] を上回る 83%の適合率でイベント情報を抽出可能であることを評価実験によって明らかにした．

今後はイベント情報の抽出の適合率のさらなる改善のため，ツイート中に含まれる URL 情報や Web の検索結果と，ツイートから抽出されたイベント情報を自動的に照合する仕組みを検討する．また機械学習だけでなく人手によるルールベースを組み合わせた手法や，抽出した自然言語の開催期間を構造化する仕組みについても検討する．また花火大会が多くなる 8 月や，イルミネーション等のイベントが多くなる 12 月といった季節の変化の影響についての評価も実施する．さらに名寄せ処理によって同一のイベント情報として判定されたイベント名称のうち，どのイベント名称が正式な名称であるか判定する仕組みについても検討する．

参考文献

[1] 内閣官房：地理空間情報基本計画，入手先 ⟨http://www.cas.go.jp/jp/gaiyou/index.html⟩ （参照 2015-07-24）.

[2] Ryong, L. and Kazutoshi, S.: Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection, *Proc. 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pp.1–10 (2010).

[3] Ling, C. and Abhishenk, R.: Event Detection from Flickr Data through Wavelet-based Spatial Analysis, *Proc. 18th ACM Conference on Information and Knowledge Management* (2009).

[4] Takeshi. S, Makoto. O, and Matsuo, Y.: Earthquake shake Twitter users: real-time event detection by social sensors, *Proc. 19th International Conference on World Wide Web* (2010).

[5] 渡辺一史，大知正直，岡部　誠，尾内理紀夫：Twitter を用いた実世界ローカルイベント抽出，第 4 回楽天研究開発シンポジウム (2011).

[6] Corinna, C. and Vladimir, V.: Support-Vector Networks, *Machine Learning*, Vol.20, pp.273–297 (1995).

[7] John, L., Andrew, M. and Feramdo, C.: Conditional randomfields: Probablistic models for segmenting and labeling sequence data, *Proc. 18th International Conference on Machine Learning*, pp.282–289 (2001).

[8] Maximilian, W. and Michael, K.: Geo-spatial Event Detection in the Twitter Stream, *Advances in Information Retrieval*, pp.356–367 (2013).

[9] Jianshu, W., Yuxia, Y., Erwin, L. and Francis, L.: Event Detection in Twitter, *ICWSM*, pp.401–408 (2011).

[10] 岡本昌之，菊地匡晃：ブログからの地域イベント情報抽出，情報処理，Vol.51, No.1, pp.14–17 (2010).

[11] 一円真治，梶　克彦，河口信夫：POI 情報統合プラットフォームの提案，マルチメディア，分散，協調とモバイル（DICOMO2013）シンポジウム，pp.1405–1412 (2013).

[12] 荒川　豊，Tagjana, S., Stephan, B., Andreas, D.：ソーシャル観光マップ—ソーシャルデータからの観光スポット抽出，マルチメディア，分散，協調とモバイル（DICOMO2013）シンポジウム，pp.1123–1132 (2013).

[13] Jaro, M.: Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, *Journal of the American Statistical Association*, Vol.85, No.406, pp.414–420 (1989).

[14] 落合桂一，鳥居大祐：時間変化する特徴語によるマイクロブログ地名曖昧性解消，情報処理学会論文誌　データベース，Vol.7, No.2, pp.51–60 (2014).

[15] Jeffely, F.: *Mastering Regular Expressions*, Oreilly & Associates Inc. (2006).

[16] 高村大也，奥村　学：言語処理のための機械学習入門（自然言語処理入門シリーズ），コロナ社 (2010).

[17] Takechi, F. and Shinichiro, T.: Japanese morphological analyzer using word co-occurrence: JTAG, *Proc. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Vol.1 (1998).

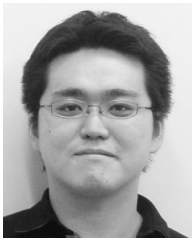
[18] Melamed, I.D.: Bitext Maps and Alignment via Pattern Recognition, *Computational Linguistics*, Vol.25, pp.107–30 (1999).

山田 渉 （正会員）

2010 年東京理科大学理工学部経営工学科卒業．2012 年東京大学大学院学際情報学府総合分析情報学コース修了．同年株式会社 NTT ドコモ入社．SNS および位置情報データ解析，ユーザインタフェースに関する研究開発に従事．ACM 会員．

菊地 悠 （正会員）

2000 年東京大学工学部精密機械工学科卒業．2002 年同大学大学院博士前期課程修了．同年株式会社 NTT ドコモ入社．SNS および位置情報データ解析の研究開発に従事．



落合 桂一 （正会員）

2006 年千葉大学工学部情報画像工学科卒業．2008 年同大学大学院博士前期課程修了．同年株式会社 NTT ドコモ入社．2014 年東京大学大学院工学系研究科技術経営戦略学専攻博士後期課程入学．SNS および位置情報データ解析の研究開発に従事．日本データベース学会会員．

鳥居 大祐 （正会員）

2001 年京都大学工学部情報学科卒業．2006 年同大学大学院社会情報学専攻にて博士（情報学）を取得．現在，株式会社 NTT ドコモにて，データマイニング，検索，リアルタイム処理，位置情報解析，機械翻訳に取り組む．

稲村 浩 （正会員）

NTT ドコモ先進技術研究所勤務．1990 年慶應義塾大学大学院理工学研究科修士課程修了．同年日本電信電話（株）入社．1994～1995 年カーネギーメロン大学計算機科学科にて訪問研究員．1998 年より NTT ドコモ．2010 年慶應義塾大学大学院開放環境科学専攻後期博士課程単位取得退学．同大学博士（工学）．モバイル環境におけるシステムソフトウェア，トランスポートプロトコル，ユーザインタフェースに関する研究開発に従事．電子情報通信学会，ACM，IEEE 各会員．本会シニア会員．

太田 賢 （正会員）

1998 年静岡大学大学院博士課程修了．博士（工学）．1999 年 NTT 移動通信網（株）入社．現在，NTT ドコモ先進技術研究所勤務．モバイルコンピューティング，端末セキュリティ，分散システムに関する研究に従事．訳書『コンピュータネットワーク第 5 版』等．電子情報通信学会会員．本会シニア会員．