

ソーシャルセンシングによる 能動的な地域密着イベント情報抽出手法の検討

Extracting Social Event Information by Active Asking to Users in Social Network Service

佐藤 圭^{1*} 池田 圭佑¹ 坂井 栞¹ 伊藤 千輝¹ 栗原 聡¹
Kei Sato¹ Keisuke Ikeda¹ Shiori Sakai¹ Kazuki Ikeda¹ Satoshi Kurihara¹

¹ 電気通信大学

¹ The University of Electro-Communications

Abstract: インターネットには神社のお祭りやアーティストのインディーズライブといった様々な地域密着イベント情報が溢れている。人々はそれらの膨大な情報を検索し入手することができる。しかし、公式にインターネット上に公開される地域密着イベント情報はあまり多くはない。これはイベントの主催者がイベント情報をインターネットに公開するための知識を持っていないからである。このためユーザはそれらのイベント情報を十分に入手することができない。この問題を解決する例として、地域密着イベント情報の集約と補完を行うサービスがある。しかし、このサービスの方式であっても、地域密着イベント情報を十分に抽出することができない問題が残されている。本稿では、地域密着イベント情報サイトが抱えるこの問題を解決するために、ソーシャルセンシングを用いたイベント情報抽出システムを提案する。このシステムでは、ソーシャルネットワークサービスのユーザに対して能動的な聴取を行い、可能な限り多種多様なイベント情報を抽出することを可能とする。

1 はじめに

インターネットには神社のお祭りやアーティストのインディーズライブといった地域密着イベント情報が多くある。近隣の人々がそのような地域密着イベントに参加すれば、その街は活性化し大きな経済効果を生み出すだろう。このように地域密着イベント情報が持つ影響力はとても大きい。観光庁 [1] によると、「地方の魅力発信プロジェクト」などの地域密着イベントを通じて観光市場を拡大させることが重要であるとしている。

インターネットの発展により、人々は多種多様な情報を検索することができるようになった。しかし、多くの情報がインターネット上に分散しているために、ユーザにとって情報を手軽に手に入れることは容易ではない。また、イベントの開催者が情報を発信する適切な方法を知らない可能性もある。例えば、イベントの開催者が開催時間や開催場所といった詳細な情報を発信してくれないと、ユーザはその情報を探ることが困難になってしまう。

そのような問題を解決しようとするサービスはいくつかあるが、我々はその中でも、地域密着イベント情報

発信サイトである『びもーる [2]』に着目している。びもーるは札幌市を中心として稼働しており、地域密着イベント情報の収集と補完を行っている。しかし、現状のびもーるにおいても多種多様な情報を抽出することができない問題がある。

そこで、我々はさらに多くの地域密着イベント情報を抽出しこの問題を解決するシステムを提案する。我々の提案するシステムでは Twitter に対してソーシャルセンシングを適用する。ソーシャルセンシングとは、ソーシャルメディアのユーザの投稿を物理センサーのように扱い情報を抽出する手法である。今日では多くの日本人が Twitter を利用しているので手軽に多くの情報を集められると考えられる。総務省 [3] によると、ソーシャルメディアは日本の社会基盤になりつつあり、日本の Twitter ユーザは全体の 28.7% であるとしている。

2 びもーるについて

『びもーる』は札幌のグルメやショッピングなどのイベントに焦点をあて毎月 1000 件以上のイベント情報を発信しているサイトである。それらの情報はウェブサイトや E メール、Twitter を通じて発信されている。びもーるの収集と発信をするシステムは小野らの提案

*連絡先： 電気通信大学大学院情報理工学研究科
〒182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: keisato@ni.is.uec.ac.jp

する「興味解析エンジン」[4]を核に構成されており、膨大な情報源からそのユーザの閲覧履歴や興味に応じて情報の発信を行うことができる。

びもーるがイベント情報の抽出からユーザに発信するまでの3つのユニットを以下に示す。

ユニット1. イベント情報抽出ユニット

このユニットでは、ウェブサイト公開された情報を2つのコレクタを用いて抽出する。1つ目のコレクタは、地元紙やイベント主催者のホームページなどのWebページに公開されている情報よりイベント情報の抽出を行う。2つ目のコレクタは、新年会やクリスマスパーティといったシーゾナルイベントの情報の抽出を行う。

ユニット2. イベント情報格納ユニット

このユニットでは、抽出したイベント情報の補完とデータベースへの登録を行う。イベント情報の補完は含まれる情報が十分でないときに行うものである。

びもーるの記事はびもーるのスタッフが2つのコレクターを用いて抽出した情報を元に記事を書いている。ここで、記事を書くのに十分な情報が含まれていない場合やWeb上に情報がない場合は、電子メールや電話などでそのイベントの主催者に問い合わせをし情報を補完する。最終的にその記事はびもーるの管理スタッフにより承認されてからデータベースへ登録される。

ユニット3. イベント情報推薦ユニット

このユニットでは、ユーザへイベント情報の提供を行う。データベースに格納されたイベント情報は、そのイベントが開催される1週間前になると情報推薦ユニットによりユーザに提供される。そのイベントが終了したらイベント情報はデータベースから削除される。

しかし、びもーるのシステムをもってしても抽出することができないイベント情報は多く存在する。これらのイベント情報がユーザに届けられることはない。

びもーるはイベント情報の抽出において、多種多様な情報を多く抽出することができないという問題を抱えている。びもーるの現在のイベント情報源はスタッフが選別を行った地方紙やウェブサイトに限られているために、抽出できる情報も限定的なものになっている。サービスを拡大していくためには、ユーザの興味を惹く様々な情報を提供していかなければならない。そのためには多種多様な情報を多く抽出する必要がある。

3 関連研究

Asurら[6]はソーシャルメディアに対して“films”でキーワード検索を行い、その投稿から統計的に映画の興行収入を予測する研究をした。しかし、統計処理を行うには多くの情報が必要であり、地域密着イベントを対象にしたびもーるではこの手法は適していないと考えられる。

Twitterからのイベント情報の抽出に注目すると、Leeら[7]のジオタグを用いた非日常地域イベントの抽出が挙げられる。Leeらの研究ではツイートに付加されているジオタグからイベント情報の抽出を行っている。しかし、この手法は人々が多く賑わっているイベントでしか有効でないことに加え、イベント情報の抽出ができるころには既にそのイベントが終わってしまっている。びもーるのサービスは、事前にイベント情報を抽出しユーザにその情報を届けることが目的であるのでジオタグを用いたイベント情報の抽出は適していない。

ここで本研究でも取り上げた榊ら[8]のソーシャルセンサーに着目する。榊らはTwitterのユーザをセンサーとするソーシャルセンシングにより現実世界の地震のデータを抽出した。榊らはソーシャルセンサーは物理センサーと同等に利用できるとしている。

同様にソーシャルセンシングを利用した研究はとして、榊ら[9]の道路の混雑状況の抽出、Nguyenら[10]のトレンドトピックの推測、Huang[11]らのインフルエンザの流行の推測が挙げられる。彼らはTwitter上でキーワード検索を用いたソーシャルセンサーにより受動的な情報抽出を行っている。しかし、本研究では受動的な情報抽出に加えて能動的な聴取を行うことでさらに情報の精度の向上を目指す。

4 提案システム

本稿ではびもーるの抱える問題に焦点を置き、Twitterに対してソーシャルセンシングを用いた地域密着イベント情報抽出を行う新しいシステムの構築を行う。

情報抽出をする上で、我々は人々の情報抽出行動に着目した。人々が情報を手に入れようとした場合、イベント情報に関するキーワードでツイートの検索を行い目的の情報を探し始めるだろう。しかし、イベントに関するツイートが少ない場合や詳細情報が十分に含まれていない場合があるかもしれない。ツイートに含まれる情報が欠けていたときには、人々はそのツイートをしたユーザに対して質問を行い、結果的に詳細情報を手に入れることができる。我々はこのような人間の行動を模倣したシステムの構築を行う。

我々の提案するシステムでは、イベント情報に関するキーワードを用いて検索を行いイベント情報ツイー

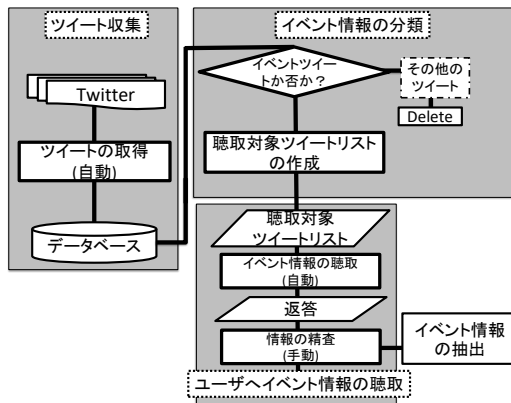


図 1: 提案システムの 3 つのユニット

トの収集を行う。イベント情報ツイートに対しては聴取を行いそのユーザからさらに詳しい情報を聴き出す。システムはユーザからの返答を元にイベント情報の補完を行い、最終的にびもーの記事データベースへ格納する。

以上より、我々は、1) ツイート収集ユニット、2) イベント情報分類ユニット、3) イベント情報聴取ユニット、の 3 つのユニットから成るシステムを提案する。システムの全体図を図 1 に示す。

4.1 ツイート収集ユニット

まずツイートの収集手法について述べる。びもーは札幌市を中心に稼働しているため、札幌市を中心としたイベント情報を集める必要がある。我々はびもーの Twitter アカウントのフォロワー約 1000 人の所在地を分類する予備実験を行ったところ、約 90% は札幌市在中だった。よって、ツイート収集はびもーのフォロワー約 8000 人を対象に REST API¹ を用いて行う。

このシステムでは収集したツイートをツイートデータとプロフィールデータに分割し MySQL データベースへ格納する。

4.2 イベント情報分類ユニット

次にイベント情報の分類手法について述べる。

提案システムでは最初にイベントツイートとそれ以外のツイートで分類を行う。

Twitter は 1 ツイートあたり 140 文字の制限さえ遵守すれば特別な制限は存在しない。しかしツイートに含まれるイベント情報は様々なフォーマットで記述されており、その状態でイベント情報の分類を行うこと

¹REST APIs - Twitter Developers <https://dev.twitter.com/rest/public>

は容易ではない。榊ら [5] は、イベント情報は“開催日時”，“開催場所”，“イベント名”のイベント三要素より構成されるとし、それを元にツイートの分類とイベント情報の抽出を行った。彼らの手法はヒューリスティックなパターンマッチングを用いたイベント三要素によるイベント情報抽出手法を提案した。しかし、榊らの研究では、『るるぶトラベル²』というサイトから、そのサイトの“イベント名”のフォーマットに沿った情報抽出を行っている。従って、榊らのイベント三要素をそのまま Twitter のイベント情報には適用することができない。

よって、我々は榊らのイベント三要素を拡張し、Twitter により適応した“イベント名”の条件付けを行う。以下に我々の提案システムで用いるイベント三要素を示す。

開催日時 正規表現により分類を行う。

開催場所 びもーのフォロワーを用いることにより札幌に限定する。

イベント名 イベント名のキーワードを「開催」「イベント」「ライブ」とし分類を行う。

システムは分類したツイートを“聴取対象ツイートリスト”へ格納する。このリストを元にシステムはユーザへ聴取を行う。

次に、“聴取対象ツイートリスト”の作成方法について述べる。聴取対象ツイートリストの作成手順を以下に示す。

ステップ 1 聴取対象ツイートリストより前日のツイートを取り出す。これはキーワードを含むか日付の正規表現に当てはまるものに限定する。

ステップ 2 同じユーザから行われたツイートを削除する。

ステップ 3 過去 7 日間で既にリストに格納されたユーザを削除する。

ステップ 4 リツイート数の多い順にソートする。

ステップ 5 上位 30 個³のイベントツイートを聴取対象ツイートリストに格納する。

次章では、イベント三要素による分類と機械学習による分類の比較実験を行う。システムの稼働実験ではより良い手法を用いる。

²るるぶトラベル <http://rurubu.travel/>

³これは API 制限を避けるためのツイート数である。

4.3 イベント情報聴取ユニット

最後にユーザへの聴取手法について述べる。このユニットでは、聴取対象ツイートリストを元に聴取を行う。ここで我々は、1) リプライを用いた聴取手法、2) 情報入力ページへのリンクより聴取を行う手法、3) ダイレクトメッセージを通じて聴取を行う手法、の3つの手法を提案する。それぞれの手法について以下で説明を行う。

リプライを用いた聴取手法

この手法ではリプライ機能を用いて詳細なイベント情報の聴取を行う。システムはリプライ文に対して挨拶文と質問文を付加する。質問文には、1. イベント名、2. 日時、3. 場所、4. その他(主催者、HP等)を含める。

実際のリプライ文の例は『突然のリプライ失礼します。』+「あなたのイベントツイートに反応! 週末のイベントを教えてください1, イベント名2, 日時3, 場所4, その他(主催者, HP等)』となる。

この質問文を埋めてもらうことで自由記述に比べイベント情報の抽出が容易になる。

情報入力ページへのリンクより聴取を行う手法

この手法では、情報入力ページへのリンクよりユーザを別ページに誘導し、そのページに入力してもらうことによりデータベースに直接イベント情報を格納する。情報入力ページへのリンクはリプライを用いてユーザへ送信する。

ダイレクトメッセージを通じて聴取を行う手法

この手法ではダイレクトメッセージを用いて聴取を行う。ダイレクトメッセージとはユーザにプライベートなメッセージを送信する機能である。この機能の文字制限は10,000文字なため、詳細なイベント情報を聴取するのに最も適していると考えられる。

4.4 システムの全体構成

我々の提案システムは上記3つのユニットより成り立っている。この節ではそれぞれのユニットの稼働時間について述べる。

我々はシステムの稼働時間のフローを、ツイートの取得を行う時間と聴取を行う時間に分けて検討をする。システムは可能な限り多くのイベント情報を抽出するために常にツイートの収集を行う。

聴取を行う時間に関して、我々はTwitterを見ているユーザが少ない時間帯に聴取を行うことは効果がない

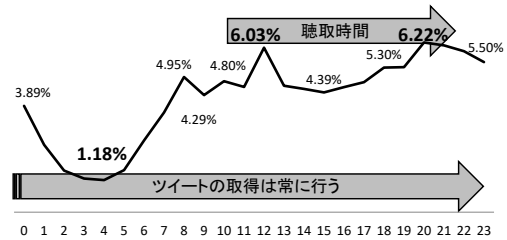


図 2: 提案システムの稼働時間

と考えた。よって、聴取を行うのに適切な時間を選択することが重要となる。図2にTwitterの利用率を推定する予備実験の結果を示す。これはびもーのフォロワーの1時間ごとのツイート数を表している。予備実験の結果より、正午12時と20時以降がツイート数が多いことが分かった。この時間帯であればユーザが反応しやすいと考えられる。よって、提案システムの聴取時間を正午12時から13時の間と18時以降に設定する。

聴取対象ツイートの作成は11時にリストの作成を行う。これは最初の聴取時間が正午12時だからである。

5 実験

この章では、我々の提案システムの有効性を確認するために、イベント情報分類精度の評価、聴取手法の評価、およびシステム全体の稼働実験の3つの実験について述べる。さらに、ここでは実験のためのデータセット作成手順についても述べる。

5.1 データセットの作成

この節ではデータセット作成手順について述べる。このデータセットはイベント情報分類精度の評価、聴取手法の評価の2つの実験で用いる。我々はびもーの約8000人のフォロワーからREST APIを用いたクローラーでツイートの取得を行った。しかし、REST APIは15分に180回までしかアクセスできない制限があるため、約8000人のフォロワー全てのツイートを1つのクローラーで取得することができない。そこで、3つのクローラーを使うことで全てのフォロワーのツイートを取得する。この3つのクローラーは24時間常時動かし続ける。

5.2 イベント情報分類精度の評価

5.2.1 実験設定

この節では、イベント三要素を用いた分類と機械学習による分類の比較実験を行うことで、我々の提案システムの評価を行う。機械学習による分類では機械学習ソフトウェアの1つである“4”を用いて、1) 決定木、2) ナイーブベイズ分類器、3) サポートベクターマシン(SVM)の3つの手法を比較する。

最初に、この実験で用いたツイートデータについて述べる。イベント情報の分類を行うために、我々は「イベント情報ツイート」とそれ以外のツイートに分類しデータセットを作成した。このデータセットには2013ツイートが含まれており、3人の学生⁵のそれぞれの判断により手作業で分類した。分類の結果、383ツイートがイベント情報ツイートであり、1630ツイートがそれ以外のツイートであった。

次にツイートデータの事前処理について述べる。我々はツイートデータを“Bag-of-Wordsモデル”を用いて成形を行った。このモデルは、テキストの並び順を無視した単語の集合と考え、単語が文書内にどこに出ているかは考慮しない表現方法である。ツイートデータをBag-of-Wordsとして扱うためには形態素解析を行い、単語ごとに分ける必要がある。そこで、本研究では形態素解析を行うためにMeCab⁶を利用する。そして、単語の出現頻度上位200件を用いて単語文書行列を作成し機械学習による分類の属性とする。以下に、単語文書行列の作成手順を示す。

ステップ1 データベースよりイベント情報ツイートを取得しそれぞれのツイートの形態素解析を行う。

ステップ2

形態素解析の結果より、各ツイートのキーワード(名詞と動詞)の集合リスト $W = \{w_n\}$ を作成する。

ステップ3

各文章におけるキーワード w_n の出現回数 w_{di} をカウントし文書ベクトル $w_d = \{c_{d1}, c_{d2}, \dots, c_{d\#(W)}\}$ を得る。

作成されたデータセットを教師データとして用いてイベント三要素による分類と機械学習による分類を行い、それぞれの手法の分類精度の評価を行う。

表 1: 分類結果

	イベント三要素	決定木
適合率	0.845	0.829
再現率	0.854	0.846
F 値	0.828	0.824
	ナイーブベイズ分類器	SVM(poly kernel)
適合率	0.81	0.842
再現率	0.814	0.855
F 値	0.812	0.837

5.2.2 実験結果

それぞれの手法を用いた分類実験の結果を表1に示す。分類の結果、SVMがイベント三要素よりも高い結果となった。しかし、F値の比較をするとそれぞれの手法に差は見られなかった。さらに、作成された決定木の上位項目として「開催、ライブ、イベント」が出現した。これは、イベント三要素による分類のキーワードに用いたものと同じであり、分類精度も同程度の精度であった。

機械学習による分類は多くの教師データを用意することが重要である。多くの教師データがあれば、今回の教師データでは出現しなかった新たなキーワードが出現するかも知れない。加えて、データ数が少ないと偏った学習をしてしまうかもしれない。機械学習の場合、多くの教師データを用意することが今後の課題となる。

よって、今回の我々の提案システムではイベント三要素による分類を使うこととする。

5.3 聴取手法の評価

5.3.1 実験設定

我々は前述した3つの聴取手法を用いてTwitterユーザに対してイベント情報の聴取を行い、それぞれの手法の評価を行う。

しかし、ダイレクトメッセージ通じてイベント情報の聴取を行う手法は、ユーザにスパム⁷だと思われてしまうことが多くあるため、イベント情報を入手することができなかった。さらに、もしユーザがTwitter社に対しスパムアカウント報告を行った場合、びもーのTwitterアカウントが停止されてしまう可能性もある。よって、ダイレクトメッセージを通じてのイベント情報の聴取を行う手法は不適切であるため、本稿ではこれを除外する。

それぞれの手法を用いたこの実験の期間と送信されたツイート数を以下に示す。

⁴Weka 3 <http://www.cs.waikato.ac.nz/ml/weka/>

⁵彼らは著者と同じ大学の学生である。

⁶MeCab とは日本語の形態素解析ソフトウェアの1つである。
<http://taku910.github.io/mecab/>

⁷ユーザに不快感を与えてしまうメッセージ

表 2: 聴取手法の実験結果

手法	ツイート数	インプレッション / ツイート	エンゲージメント / ツイート
リプライを用いた聴取手法	177	20.70	0.932
情報入力ページへのリンクより聴取を行う手法	225	13.56	0.556
手法	URL クリック数	詳細クリック数	リプライ数
リプライを用いた聴取手法	0	76	11
情報入力ページへのリンクより聴取を行う手法	31	24	1
手法	提供されたイベント情報の数	びもーるに掲載されたイベント情報の数	
リプライを用いた聴取手法	6	4	
情報入力ページへのリンクより聴取を行う手法	2	2	

1. 手法： リプライを用いた聴取手法

実験期間： 2015 年 10 月 27 日 - 2015 年 11 月 4 日

聴取時間帯： 12:00 - 13:00 & 18:00 以降

ツイート数： 177

2. 手法： 情報入力ページへのリンクより聴取を行う手法

実験期間： 2015 年 10 月 8 日 - 2015 年 10 月 22 日

聴取時間帯： 12:00 - 13:00 & 18:00 以降

ツイート数： 225

5.3.2 実験結果

表 2 にそれぞれの聴取手法の結果を示す。

“インプレッション”はそのツイートがユーザに読まれた回数を示している。“エンゲージメント”はそのツイートがユーザのタイムラインに表示された回数を示している。それぞれの手法でツイート数が違うため、これらの値は各件数で割った 1 ツイートあたりの平均で比較する。“URL クリック数”はツイートに張られたリンクのクリック数を表している。“詳細クリック数”は、インプレッションとは異なり、ユーザが実際にツイート情報をクリックし詳細を確認した数を表している。

表 2 より、インプレッション、エンゲージメント、詳細クリック数がリプライを用いた聴取手法の方が良い結果となっていることがわかる。

よって、我々の提案システムではリプライを用いた聴取手法を採用する。

5.4 システムの稼働実験

イベント三要素によるイベント情報の分類とリプライを用いた聴取手法によりシステムの稼働実験を行う。この実験の概要を以下に示す。

システムの稼働実験の概要：

分類手法： イベント三要素による分類

聴取手法： リプライを用いた聴取手法

実験期間： 2015 年 11 月 10 日 - 2016 年 1 月 29 日

聴取時間帯： 12:00 - 13:00 & 18:00 以降

ツイート数： 1755

この実験では、システムがびもーるのフォロワーのツイートを収集し、イベント三要素による分類を行った。分類を行ったツイートを元に“聴取対象ツイートリスト”を作成し、リプライを用いた聴取手法により聴取を行った。実験の結果を表 3 に示す。ここでユーザに送られたリプライの数は 1755 ツイートであり、98 個のリプライが返ってきた。このうちイベント情報を含むものは 71 ツイートであり、実際にびもーるに掲載された数は 26 個である。

これよりこの実験の結果について考察を行う。

この実験では、提供されたリプライよりもびもーるに掲載された情報の方が少なくなりました。これはびもーるの基幹システムの仕様からなる問題である。

表 3: システムの稼働実験の結果

手法	ツイート数	インプレッション /ツイート	エンゲージメント /ツイート
リプライを用いた手法	1755	23.33	1.088
手法	URL クリック数	詳細クリック数	リプライ数
リプライを用いた手法	0	1138	98
手法	提供された イベント情報の数	びもーるに掲載された イベント情報の数	
リプライを用いた手法	71	26	

びもーるの基幹システムは、あるイベントが開催される1週間前でないとそのイベントの情報を掲載することができない。しかし、提供されたイベント情報は開催が直前に迫っているものが多く占めていた。このために、71個のイベントツイートが提供されたにもかかわらず実際に掲載された情報は半分以下の26個となってしまった。びもーるの基幹システムを対応させこの問題を解決できれば倍以上の情報が掲載できるようになるだろう。

次に、重複したユーザに送られたツイート数に関して述べる。我々の提案システムでは「イベント情報分類ユニット」において重複したユーザに聴取を行わないようにしていた。しかしこの実験では1755ツイートが送られたうち1036ツイートが重複したユーザ宛のツイートであった。同じユーザに何回も聴取を行ってしまい、そのユーザに不快感を与えてしまった例もあった。よって、重複したユーザに聴取を行うことは好ましくないことであるため、聴取対象ツイートのリストの作成の際に重複したユーザを含まないようにする必要がある。一方で「イベント情報をゲットしたらツイートします!」といった返信を提供していただいた聴取に前向きなユーザがいることも分かった。このようなユーザには積極的な聴取を行うことでイベント情報を提供してもらえようと考えられる。

続けて、リプライ文のフォーマットについて述べる。この実験では、我々が返信文のフォーマットを用意したにも関わらず、そのフォーマット通りに回答をしていただいたユーザはわずかであり、多くのユーザは自由記述により情報提供をしていただいた。さらに、何人かのユーザからはイベント情報が掲載された公式サイト URL や、他のユーザを見るとイベント情報が掲載されているといった情報が提供された。よって、ユーザから多くの情報を聞き出すための聴取文を検討する必要がある。また、URL が提供された際の Web スクレイピングの実装もする必要があるだろう。

最後に、イベント情報の開催場所について述べる。我々はこの実験で、札幌のイベント情報の抽出を行うためにびもーるのフォロワーに対して聴取を行った。提供されたイベント情報の開催場所を表4に示す。表4

表 4: イベント情報の開催場所

都道府県 市町村	北海道 札幌市	北海道 その他	その他	合計
イベント情報の数	51	10	10	71

は札幌市のイベント情報が約70%を占めていることを示している。その一方で、『北海道展』のような北海道以外で開催されるイベント情報も提供された。札幌のイベント情報の抽出のためにびもーるのフォロワーに対して聴取を行うのは有効であると考えられる。しかし、イベントの場所を札幌周辺に限定するために、聴取文を再検討する必要がある。

6 今後の展望

3つの実験を通じて、1) イベント分類手法に機械学習を用いる場合の教師データの数の少なさ、2) 提供されたイベント情報よりもびもーるに掲載できた情報の数が少なくなってしまう、3) 重複したユーザに何度も聴取を行ってしまった、4) こちらはフォーマットを指定したがユーザは自由記述により回答をした、5) 札幌周辺以外のイベント情報が提供された、の5つの課題が生まれた。

1) 教師データの少なさはシステムを長い期間稼働させることで対応できると考えられる。

2) びもーるに掲載できた情報が少なくなってしまう問題はびもーるの基幹システムの仕様を改訂することで対応可能であるだろう。

3) 重複したユーザに何度も聴取を行ってしまった問題を解決するためには、聴取対象ツイートリスト作成手順を見直す必要がある。今回の実験では、聴取に対して前向きなユーザと不快感を与えてしまったユーザの2種類のユーザがいることが分かった。我々はユーザからの返答に対して感情推定を行い聴取対象ツイートリスト作成へ重み付けを行うことで、より効率的にイベント情報の聴取が行えると考えている。加藤ら[12]は、Twitterのツイートをユーザ名、ハッシュタグ、URL、カッコ付き文字、名詞、動詞、形容詞、形容動詞の要素

に分解し感情情報を付加を行っている。感情情報の付加では、中村 [13] の感情表現辞典を元に加藤らが定義した絶対感情語を元にしてツイートの感情値の推定を行っている。我々の提案システムにおいても、ユーザからの返答を考慮した感情語を定義しそれぞれのユーザの特性を分類することで、聴取対象ツイートのリスト作成への重み付けに活かすことができるのではないかと考えている。

4) ユーザが自由記述により回答をした, 5) 札幌周辺以外のイベント情報が提供された, の2つの課題はユーザへの聴取文を再検討することで対応可能であると考えている。また, ユーザから URL の情報が提供された場合には Web スクレイピング技術を用いた実装をすることでイベント情報の抽出を行うことができるだろう。これらの課題を解決することでさらに多種多様なイベント情報を抽出できるようになりより良いサービス展開が期待できる。

7 おわりに

本稿では地域密着イベント情報発信サイトである『びもーる』に着目し問題の洗い出しを行った。さらに, その問題を解決するためにソーシャルセンシングを適用したシステムを提案した。このシステムでは Twitter のツイートからイベント情報の抽出を行い, そのツイートをしたユーザに対し聴取を行うことで情報の補完を行った。

提案システムは 1) ツイート収集ユニット, 2) イベント情報分類ユニット, 3) イベント情報聴取ユニット, の3つにユニットから成り立っている。本稿では, 2) のユニットのイベント情報分類手法の比較実験, 3) のユニットのイベント情報聴取手法の比較実験に加え, それぞれの実験で良い結果であった手法を用いてシステム全体の稼働実験を行った。

イベント情報分類手法の比較実験では“イベント三要素による分類”と機械学習による分類の比較実験を行った。イベント情報聴取手法の比較実験では, 1) リプライを用いた聴取手法, 2) 情報入力ページへのリンクより聴取を行う手法, 3) ダイレクトメッセージを通じて聴取を行う手法, の3つの手法の比較実験を行った。システムの全体の稼働実験では“イベント三要素による分類手法”と“リプライを用いた聴取手法”を用いた。

それぞれの実験を通じて, 1) イベント情報分類手法に機械学習を用いる場合の教師データの数の少なさ, 2) 提供されたイベント情報よりもびもーるに掲載できた情報の数が少なくなってしまう, 3) 重複したユーザに何度も聴取を行ってしまった, 4) こちらはフォーマットを指定したがユーザは自由記述により回答をした, 5)

札幌周辺以外のイベント情報が提供された, の5つの課題があることが分かった。これらの課題は, びもーるの基幹システムの修正, 感情推定による聴取対象ツイートへの重み付け, Web スクレイピング技術の実装により解決できると考えられる。これにより多種多様なイベント情報の抽出が可能になり, より良いサービス展開へ繋げることができるだろう。

参考文献

- [1] 観光庁:平成28年版観光白書, <http://www.mlit.go.jp/statistics/file000008.html> (2016)
- [2] あなた情報マガジンびもーる, <http://bemall.jp/>
- [3] 総務省:平成28年度版情報通信白書, [http://www.soumu.go.jp/johotsusintokei/whitepaper/\(2016\)](http://www.soumu.go.jp/johotsusintokei/whitepaper/(2016))
- [4] 小野 良太, 山下 晃弘, 川村 秀憲, 鈴木 恵二: イベント開催情報推薦のためのスコアリングの検討, 観光と情報: 観光情報学会誌, Vol.11, No.1, pp.23-34(2015)
- [5] 榊 剛史, 那須野 薫, 柳原 正: 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, 人工知能学会全国大会論文集, Vol.27, pp.1-4(2013)
- [6] Asur, Sitaram and Huberman, Bernardo A.: Predicting the Future with Social Media, *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol.1 pp. 492-499(2010)
- [7] Lee, Ryong and Wakamiya, Shoko and Sumiya, Kazutoshi: Discovery of unusual regional social activities using geo-tagged microblogs, *World Wide Web*, Vol.14, No.4, pp.321-349(2011)
- [8] Sakaki, Takeshi and Okazaki, Makoto and Matsuo, Yutaka: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, *Proceedings of the 19th International Conference on World Wide Web '10*, pp851-860(2010)
- [9] T. Sakaki and Y. Matsuo and T. Yanagihara and N. P. Chandrasiri and K. Nawa: Real-time event extraction for driving information from social sensors, *2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp221-226(2012)
- [10] Duc T. Nguyen and Jai E. Jung: Privacy-Preserving Discovery of Topic-Based Events from Social Sensor Signals, *The Scientific World Journal Volume 2014*
- [11] J. Huang and H. Zhao and J. Zhang: Detecting Flu Transmission by Social Sensor in China, *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, pp/1242-1247(2013)
- [12] 加藤, 慎一朗 and 濱川, 礼: Twitter から得られる自然言語情報を用いて行う単語への感情付加手法, 研究報告 ヒューマンコンピュータインタラクション (HCI), Vol.2012-HCI-148, No.16, pp.1-8(2012)
- [13] 中村, 明: 感情表現辞典, 東京堂出版 (1993)