

ソーシャルセンシングを用いた 能動的な地域密着イベント情報の抽出

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^{*2} 栗原 聡^{*1}
Kei Sato Keisuke Ikeda Shiori Sakai Kazuki Ikeda Satoshi Kurihara

^{*1}電気通信大学

^{*2}ネットワンシステムズ株式会社

^{*1} The University of Electro-Communications ^{*2} Net One Systems Co., Ltd.

There are a lot of local-event information, such as shrine festivals and any kind of concert on the Internet. We are able to search for a such large amount of varied information. However, there is not much local-event information officially published in the Internet. Event organizers do not have sufficient knowledge on how to announce such information using the Internet. For this reason, users cannot get local-event information enough. There is the services to resolve these problems. We are focusing on the local event information sites as one of the service. Still, there are times when the sites cannot extract local event information there enough. In this paper, we propose an event-information-extraction system that uses social sensing. With this system, we actively ask social networking service users to obtain detailed information about events as well as extract a wide variety of information.

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が多くある。このような地域密着イベント情報を通じて観光市場を拡大させることが重要である。

しかし、イベントの主催者がそのようなイベント情報を発信しようとした際に十分なノウハウを持っていないために、情報が十分ではないままインターネット上に分散されてしまうことがある。この場合、ユーザがこのイベント情報を検索することは難しくなってしまう。そのような問題を解決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 [1]』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。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。びもーるの収集と発信をするシステムは小野ら [2] の提案する「興味解析エンジン」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。これに対し、本研究では情報抽出では収集が間に合わないイベントを不完全な情報しか抽出できない状況において、必要な情報を持っていると推定されるユーザに能動的に聞き取りを行うことで、より正確に旬なイベント情報を収集する方式を提案する。

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。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。

2. 関連研究

Twitter からのイベント情報の抽出に注目すると、Lee ら [3] 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 [4] 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としては、榊ら [5] の道路の混雑状況の抽出、Nguyen ら [6] のトレンドトピックの推測、Huang [7] 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。

3. 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。

我々の提案するシステムでは受動的な情報の取得に加え、取得した情報の発信者に対し能動的な聴取を行うことで詳しいイベント情報の抽出を目指す。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、最終的にびもーるの記事データベースへ格納する。

このシステムは、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 つのユニットから成り立つ。システムの全体図を図 1 に示す。

連絡先: 佐藤 圭,

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

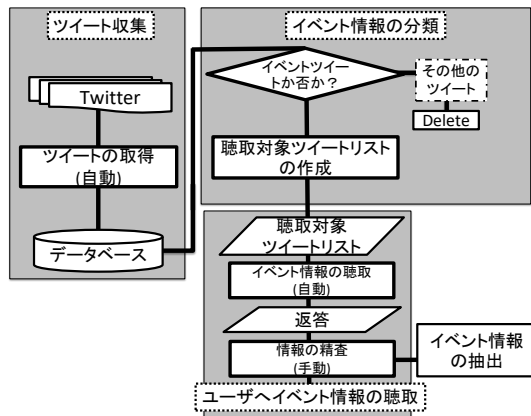


図 1: 提案システムの 3 つのユニット

それぞれのユニットについて以下に述べる。

3.1 ツイート収集ユニット

びもーは札幌市を中心に稼働しているため、ツイート収集はびもーの Twitter アカウントのフォロワー約 8000 人に対して REST API^{*1} を用いて行う。これは、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% が札幌市中であることが分かったためである。これにより札幌市のイベント情報の抽出を図る。

3.2 イベント情報分類ユニット

イベント情報分類ユニットでは、最初にイベントツイートとそれ以外のツイートの分類を行う。

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は容易ではない。榊ら [8] 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 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ーへ聴取を行う。

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

*2 るぶトラベル <http://rurubu.travel/>

聴取対象ツイートリストは、同じユーザーから行われたツイートの削除、過去 7 日間で既にリストに格納されたことのあるユーザーの削除を行うことにより、同じユーザーに重複して聴取を行わないようにする。聴取対象ツイートリストの上位 30 個のイベントツイートに対して聴取を行う。

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。

3.3 イベント情報聴取ユニット

イベント情報聴取ユニットでは、聴取対象ツイートリストを元にユーザーへの聴取を行う。

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ー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ーへ送信する。

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ー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。

3.4 システムの全体構成

ここでは上記 3 つのユニットのそれぞれ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 を見ているユーザーが少ない時間帯に聴取を行うことは効果がないと考えた。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。びもーのフォロワーの Twitter の利用率を推定する予備実験を行ったところ、正午 12 時と 20 時以降にツイート数が増えることが分かった。この時間帯であればユーザー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。

聴取対象ツイートリストの作成は最初の聴取時間正午 12 時の 1 時間前である 11 時に行う。

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

4. 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。

4.1 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。

4.2 イベント情報分類精度の評価

4.2.1 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。機械学習による分類では、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。

この実験で用いたツイートデータは、2013 ツイートを3人の学生の判断によりイベント情報ツイートとそれ以外のツイートに手作業で分類した。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。

このツイートデータを、MeCabにより形態素解析を行い、Bag-of-Wordsモデルによる成形を行った。Bag-of-Wordsモデルとは、テキストの並び順を無視した単語の集合と考え生成する、単語が文書内のどこに出てくるかは考慮しない単語文書行列のことである。この実験では単語文書行列の出現頻度上位200件を教師データとして用いる。

4.2.2 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。

4.3 聴取手法の評価

4.3.1 実験設定

我々は前述した3つの聴取手法を用いてTwitterユーザー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、ユーザーにスパムだと思われる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、もしユーザーが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実験結果を以下に示す。

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

インプレッション/ツイート：20.70

エンゲージメント/ツイート：0.932

詳細クリック数：76

提供されたイベント情報の数：6

びもーに掲載されたイベント情報の数：4

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

インプレッション/ツイート：13.56

エンゲージメント/ツイート：0.556

詳細クリック数：24

提供されたイベント情報の数：2

びもーに掲載されたイベント情報の数：2

4.3.2 実験結果

“インプレッション”はそのツイートがユーザー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ー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。“詳細クリック数”は、インプレッションとは異なり、ユーザー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。実験結果より、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ったため、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。

4.4 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。この実験では、システムがびもーのフォロワーのツイートを集集し、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。この実験の概要と結果を以下に示す。

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
- インプレッション/ツイート： 23.33
- エンゲージメント/ツイート： 1.088
- 詳細クリック数： 1138
- 提供されたイベント情報の数： 71
- びもーに掲載されたイベント情報の数： 26

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、71個のイベント情報が提供された。実際にびもーに掲載された数は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に掲載された情報の方が少なくなってしまった。これはびもーの基幹システムの仕様からなる問題である。びもーの基幹システムは、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。びもー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。一方で聴取に前向きなユーザがいることもわかった。聴取対象ツイートリスト作成の際にこのようなユーザに対する重み付けを行う必要があると考えられる。

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、我々が返答文のフォーマットを用意したにも関わらず、多くのユーザから自由記述により情報提供がされた。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や、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。

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71個のうち約70%の51件が札幌市のイベント情報であったその一方で、『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。

5. 今後の展望

我々は多種多様なイベント情報を効率的に抽出するには聴取対象ツイートリスト作成手順の再検討が第一であると考えている。システムの稼働実験を通じて重複したユーザに何度も聴取を行ってしまった結果、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[9]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[10]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。

参考文献

- [1] あなた情報マガジンびもー, <http://bemall.jp/>
- [2] 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)
- [3] Lee, Ryong and Wakamiya, Shoko and Sumiya, Kazutoshi: Discovery of unusual regional social activities using geotagged microblogs, *World Wide Web*, Vol.14, No.4, pp.321-349(2011)
- [4] 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)
- [5] 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)
- [6] Duc T. Nguyen and Jai E. Jung: Privacy-Preserving Discovery of Topic-Based Events from Social Sensor Signals, *The Scientific World Journal Volume 2014*
- [7] 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)
- [8] 榑 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)
- [9] 加藤 慎一郎 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)
- [10] 中村 明: 感情表現辞典, 東京堂出版 (1993)